Statistical assessment of repeated measures data with application in oncology

Seegobin, Seth David

Awarding institution:
King's College London

Statistical assessment of repeated measures data with application in oncology

Seth David Seegobin / 1208745

Department of Medical and Molecular Genetics

January 2017

Thesis submitted to King's College London in fulfilment of the degree of

Doctor of Philosophy

Dedication

This thesis is dedicated to the memory of my father, Dr Ronald Seegobin. You set my feet on this path and I wish you were here to see me finish. I remain inspired by your commitment to patient care. The values you instilled in me are the determination in every page.

Acknowledgements

It has been my privilege to work with Professor Cathryn Lewis. I have enjoyed the opportunity to watch and learn from her knowledge and experience; her frequent insights and patience were always appreciated. I am very proud to have been a member of your research group, thank you.

I would not have been able to progress my linear regression work without the wise counsel of Dr Irene Rebollo-Mesa, who consistently challenged me to improve my understanding. I am grateful for her guidance and support.

A special mention for the members of the Purine Research Laboratory – Dr Jeremy Sanderson, Dr Tony Marinaki and Dr Adele Corrigan – who were kind enough to share their data which made this thesis possible.

To the many members of the Statistical Genetics Group, both at Guy's Hospital and the Institute of Psychiatry, thank you for your research input. I have learned from watching all of you and benefited from your questions and ideas regarding my own work; special thanks are owed to Drs Ken Hanscombe, Mathew Traylor and Raquel Iniesta for their help with R graphs and review of sections of my thesis.

To the nights I can never remember with the people I can never forget. I am indebted to all my friends at 'Guy's Club Fridays' who have supported me over the last few years. The group is too large to name everyone, but special thanks are owed to Matthew Shannon, Chronis Kemos, Inês Barbosa, Jake Saklatvala, Dorita Tsourouktsoglou, Jon Temple and Laura Demandt. Thank you all for laughing and crying with me – usually at the right time for both. Last but not least, to my training partner, cultural attaché and friend for life - Christos Petridis. I cannot count the ways in which you helped

over the last four years. Thank you for always being there whenever I needed you. And even when I didn't think I needed help you were there, just in case.

I wish to thank my mother and brothers for all the moral support they've given me. And to my furry friends, Cassius and Isabel – you were both there at the start, we miss you every day Isabel.

Finally, to my wife, Helena. Without your love and understanding I would not have completed this work. I promise no more schooling and regular holidays from here on in.

Preface

Prior to being a student at King's College London, I worked as a clinical trials biostatistician within oncology. In analysing and reviewing the results of the trials, I was forced to recognise that, every day, patients are exposed to medications that are of limited or even no benefit to them. Classical clinical trials harvest a handful of measurements from thousands of people. These trials typically are unable to distinguish patient characteristics that interact with response within a given treatment arm, resulting in treatment decisions that follow a 'one size fits all' approach. My desire to explore the potential of genetics to explain variability in patient drug response provided the impetus for me to pursue the work presented in this thesis.

While deciding which area to focus on for my PhD research, a clinical colleague approached me with an offer of collaboration and asked me to conduct a repeated measures analysis of clinical trial data.

These data included a baseline measurement of outcome response prior to the initiation of treatment. While this was not the kind of analysis that I was used to from oncology trials, I recalled from lectures on the subject that I could include the baseline as a covariate (ANCOVA), calculate the change between baseline and each subsequent measurement (change score – CSA) or retain the baseline measurement as part of the outcome vector. I searched the internet for guidance on which of the three analysis methods to use, and found only toy examples (in which all analysis methods yielded the same result), or complex discussions about bias with the ANCOVA, and numerous references to the choice of analysis method being akin to 'Lord's Paradox' (with which I was previously unfamiliar).

I realised that it would take some time to truly know the implications of each method, however, my clinical colleague was keen to move ahead, without waiting for me to explore the statistical subtleties of each method. We therefore chose to use the analysis with the highest *p*-value for the primary endpoint. I rationalised this choice as statistical-conservatism, secure in the knowledge that I could not be accused of 'cherry-picking' the most favourable *p*-value. The paper was published in a reputable journal [1], but this was cold comfort since I was still no closer to understanding why each analysis method had yielded different results. When my supervisor asked me for a justification of my analysis choice my 'conservatism' rationale sounded hollow. With her encouragement, I set out to

understand exactly when and why different methods of handling the baseline measurement yield different results. The work resulting from this exploration forms Part 2 of the thesis (Chapters 8-11). To quote my primary supervisor, “a PhD is a training degree and one of the few opportunities to pursue your learning interests without constraint”, and in this regard my work on repeated measures linear models was an unparalleled success. From programming simulations of correlated and missing data to linear algebra, each step of this project challenged me to learn new things and expand my view of the tests that we so often use in statistics.

For a data-driven component to the thesis, my supervisor used her collaboration with the Purine Research Laboratory to secure for me data from cancer patients treated within St. Thomas Trust. These data contained baseline, characteristics, germline genetics and patient response (efficacy and safety endpoints). This data provided the platform from which I could pursue my original interest in understanding sources of variation in patient response – or as it is better known, personalised medicine.

The result is a thesis with two distinct parts. As my primary research interest is personalised medicine, the first half of my thesis (Chapters 2-6) uses patient data to identify genetic and baseline patient characteristics for their association with efficacy and safety responses. My early work forms the second half of my thesis (Chapters 7-11) which explores repeated measures linear regression.

It is my hope that this thesis has contributed to a refined understanding of repeated measures linear regression and the identification of prognostic factors associated with cancer therapy that can be used to stratify patient response.

Abstract

Statistical analysis methods that analyse complex data appropriately are necessary for medical research to fulfil its aims of improving cancer treatment. To move towards this goal, this thesis uses data from clinical studies and simulation to explore two different areas: (1) genetic associations with efficacy and safety of chemotherapy, and (2) repeated measures analysis of clinical trial data.

Several studies have reported that safety events occurring during chemotherapy are predictors of significantly longer survival for cancer patients. Recent research has identified germline genetic variants associated with either safety or efficacy drug response outcomes. We sought to understand the similarities and differences in the genetic association signals between safety and efficacy endpoints. Our work highlights the difficulties in combining cohorts of patients across cancer types, since differences between cohorts with respect to baseline characteristics and efficacy responses prohibit the use of meta-analysis for the discovery of response-associated factors. Our work confirms that baseline patient characteristics can be important prognostic factors in drug response, however, we conclude that the addition of baseline factors as covariates does not assist in the identification of genetic variants associated with response. Lastly, we develop a novel graphical method to describe the similarities in genetic association results between any two clinical endpoints measured in cancer studies.

Baseline values are commonly measured in clinical trials to help assess drug response following randomisation. Treatment effects on mean change from baseline can be analysed: 1) including the baseline value as part of the treatment response, 2) using only the post-randomisation values in the response analysis only the post-randomisation values and baseline as a covariate, or 3) using the calculated change from baseline value as the dependent variable. We consider each of these analysis methods for their accuracy and precision in estimating the between-group difference in the mean change from baseline. We conclude that the method by which the baseline responses are used in the analysis influences both the accuracy and the efficiency of identifying the response slope difference between treatment arms. The difference in accuracy and precision between methods depends on the number of post-randomisation assessments, with-patient correlation strength and correlation structure of repeated measurements.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

Abbreviation	Meaning
AA	Any adverse event
ADME	Adsorption, metabolism and excretion
AE	Adverse event. Note: 'severe adverse event'
AIC	Akaike information criterion
AML	Acute myeloid leukaemia
ANC / ANCOVA	Analysis of covariance. A repeated measures analysis in which only the post treatment administration values are analysed in the response vector and the baseline measurement is fit as a covariate independent variable.
ANOVA	Analysis of variance
AR	Autoregressive; a common covariance structure. Used interchangeably with AR(1).
ASCO	American Society of Clinical Oncology
AUC	Area under the curve
BMI	Body mass index
BP	Base pair position
BSA	Body surface area
Chr	Chromosome number – only in tables BP
CI	Confidence interval
cLDA	Constrained longitudinal data analysis
CS	Compound symmetry; a common covariance structure
CSA	Change score analysis. A repeated measures analysis utilising change from baseline scores as the response variable.
CSF	Colony stimulating factor
CTCAE	[National Cancer Institute] Common Terminology Criteria for Adverse Events v4.7– (also called Common Toxicity Criteria for Adverse Events)
DLT	Dose-limiting toxicity
ECOG	Eastern Cooperative Oncology Group. One of the largest clinical cancer research organisations in the United States.
EDTA	Ethylenediaminetetraacetic acid
EIGENSOFT	EIGENSOFT v3.0. Software for principal components analysis. Population structure analysis in genetic association.
EMA	European Medicines Agency
FDA	U.S.A. Food and Drug Administration

G-CSF	Granulocyte colony stimulating factor
GID	Gastrointestinal disorder. A system organ class term, the highest level of the MedDRA hierarchy. A composite term comprised of the specific adverse events of mucositis, constipation, diarrhoea, vomiting, and nausea.
GM-CSF	Granulocyte-macrophage colony stimulating factor
GWAS	Genome-wide association study
H_0	Null hypothesis
H_A	Alternate hypothesis
HR	Hazard ratio
LD	Linkage disequilibrium. A phenomenon whereby the frequency of two alleles at different loci is higher or lower than would be expected if the loci were independent and associated randomly.
LDA	Longitudinal data analysis
MAF	Minor/risk allele frequency
MANOVA	Multivariate analysis of variance
MedDRA	Medical Dictionary for Regulatory Activities, version 12.1
MRM	Mixed-effect regression model
NCI	National Cancer Institute
NE	Not estimable/not able to estimate
NEU	Neutropenia. A condition characterised by low level of neutrophils, which are a type of white blood cell.
NIH	National Institutes of Health
NSCLC	Non-small-cell lung cancer
OD method	Repeated measures sample size calculation based upon the work by Overall and Doyle.
OLS	Ordinary least squares
OR	Odds ratio
OS	Overall survival
PC	Principal component
PFS	Progression free survival
PH	Proportional hazards
PRS	Polygenic risk scores
PS	[ECOG] performance status
PV	Pharmacovigilance
Q-Q plot	Quantile-quantile plot

R	R software version 3.0, developed by R Core Team, 2013. Open source programming language and software for statistical computing and graphics supported by R Foundation for Statistical Computing.
RCT	Randomised controlled trials
RF	Renal function
RMA	Repeated measures analysis, which incorporates the baseline measurement in the analysis response vector.
RSID	Reference SNP cluster ID
SAS	SAS version 9.4. Statistical computing software produced by SAS Institute.
SD	Standard deviation
SNP	Single nucleotide polymorphism
SOC	System organ class. Most general level of MedDRA adverse event classification.
SST	Total sum of squares
TPMT	Thiopurine methyltransferase
TT method	T-test based sample size calculation for two independent groups assuming common variance.
TVC	Time-varying covariate
UN	Unstructured
WBC	White blood cells

Chapter 1. Introduction

> It is far more important to know what person the disease has than what disease the person has.
>
> – Hippocrates of Cos

1.1 Introduction to personalised medicine in cancer

Clinical outcomes in response to drugs exhibit large inter-patient variability in terms of treatment efficacy and drug toxicity. Patient demographics, behaviours and genetics all are underlying factors hypothesised to contribute to this variation [2]. Personalised medicine is essentially the ability to tailor treatments, as well as prevention strategies, to the unique characteristics of each person. One vision of personalised or stratified medicine is to use knowledge of these sources of variation in order to identify subgroups or strata of patients who are more (or less) likely to respond to a treatment [3].

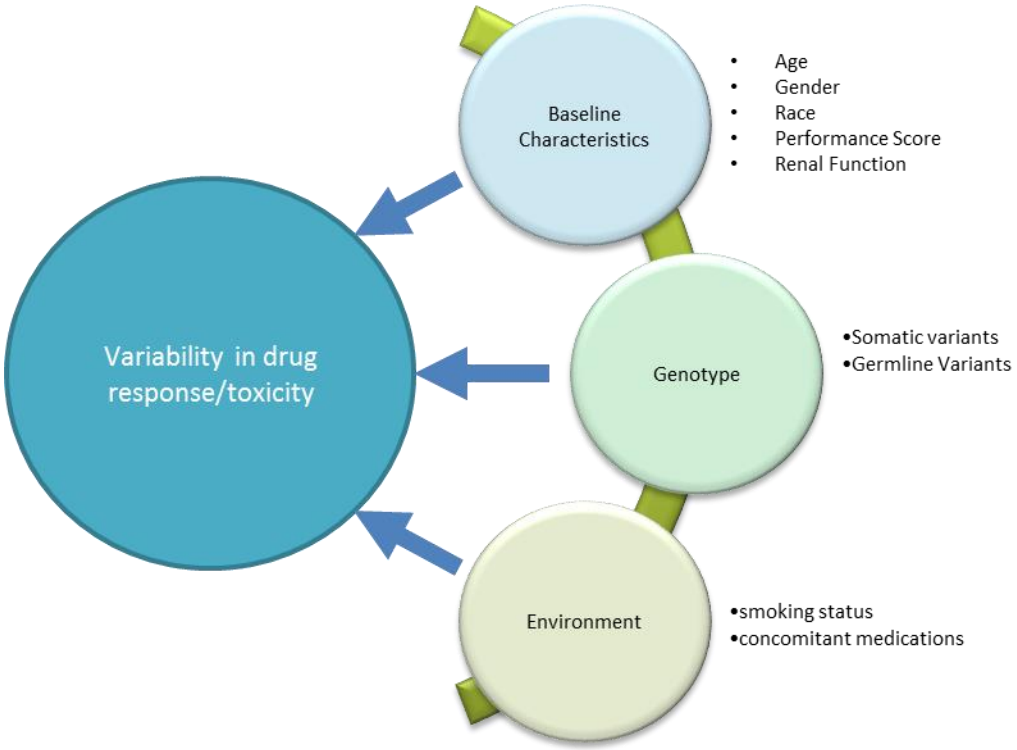


Figure 1.1.1 Sources of variability in drug response.
Variation in drug response is known to be influenced by an individual patient's genetic and environmental factors as well as baseline characteristics.

1.2 Baseline characteristics and behaviours

The concept of personalised medicine is not new. Within oncology, the concept of matching treatments to patients to achieve remission and avoid re-occurrence is well established. Clinicians have long observed that different patients respond differently to medical interventions, and it is standard clinical practice to optimise an individual's treatment based upon differences in patient characteristics or prognostic factors [4]. A prognostic factor provides information about patient outcomes irrespective of therapy; a predictive factor is one that modulates the effect of a therapeutic intervention [5].

During recent decades, numerous studies have investigated prognostic and predictive factors for lung and ovarian cancer survival. These studies have identified disease stage [6-9], performance status (PS) [7, 8, 10-13], gender [11, 14, 15] and weight loss [7, 14, 16, 17] as strong prognosticators of survival and risk of toxicity. The role of other factors, including age [7, 8, 11, 18, 19], ethnicity [19-21] and smoking status [22, 23] as significant prognostic factors in lung and ovarian cancer remains controversial. Accurate characterisation of prognostic factors is essential for predicting patient outcomes, reporting the results of comparative studies in cancer therapy, and comparing new investigational results to previous studies.

Translating the concepts of personalised medicine into direct benefit to patients requires identification of the many factors that influence patient variability in response. This creates the need for extensive characterisation of the various baseline characteristics factors that influence patient response.

1.3 Pharmacogenomics

Pharmacogenomics is the part of personalised medicine that aims to identify inherited genetic variants that may in part explain inter-patient response variation based upon genetic differences between patients. The terms pharmacogenomics and pharmacogenetics are often used interchangeably; within this thesis we use the term pharmacogenomics. Within the field of oncology, applications of pharmacogenomics cover two main areas: avoidance of safety reactions - adverse events (AE), and prediction of treatment efficacy for a given patient [24]. Some pharmacogenomic

tests are now in standard clinical practice. For example, patients are tested for presence of *HLA-B**57:01 before treatment with abacavir, as this variant increases risk of abacavir hypersensitivity. Low thiopurine methyltransferase (TPMT) enzyme activity can lead to toxic effect when azathioprine and similar drugs are taken. Pharmacogenomic tests are used to identify people with no TPMT activity (for whom a different drug should be used) or low activity (who need a reduced dosage).

To identify variants associated with a particular drug therapy, genome-wide-association-studies (GWASs) are often performed. Such studies assess single-nucleotide-polymorphism (SNP) variation between individuals and determine how this genetic variation is associated patient response [25]. Within oncology, genetic association studies have previously been used to successfully correlate particular genomic signatures with drug efficacy response and risk of toxicity development [26] in patients treated with irinotecan (colorectal cancer) [27], capecitabine (breast and colorectal cancer) [28, 29], cyclophosphamide (leukaemia) [30], and tamoxifen (breast cancer [31], metastatic colorectal cancer [32, 33], chronic myeloid leukaemia [34]).

1.3.1 Germline vs. somatic variation

In cancer pharmacogenomics, there are two genomes that may influence variation to drug response: the genome of the person with cancer (germline) and the genome of the malignant cells (somatic). Almost all cancers are associated with damaged DNA that allows cells to grow unchecked. This means that tumour genomes are a mixture of cancer and normal cells and therefore each tumour contains inherited (germline) and tumour-specific (somatic) variants.

These tumour-specific genetic changes provide potential targets for cancer treatment and have already started to transform treatment strategies in oncology. Afatinib (lung), alectinib (lung), bevacizumab (lung and ovarian), ceritinib (lung), gefitinib (lung), necitumumab (lung), nivolumab (lung), ramucirumab (lung) are all examples of targeted therapies [35] that use somatic mutation induced cellular changes to target cancer cells with greater precision than traditional chemotherapy agents.

Despite the success of targeted agents, the mechanism that provides their success is also their key limitation. Within a tumour, the cancer cells may be heterogeneous; different sections of the tumour

may be derived from different clonal expansions. Consequently, within the same patient, different tumours and even different cells within the same tumour will carry alternate somatic mutations. This means that treatment strategies based upon somatic mutations will frequently not be able to target all tumours within a patient. For this reason, targeted therapies rarely result in durable tumour regressions and most patients will ultimately experience disease progression despite an initial period of response [36].

By contrast, germline mutations are present in every tumour and non-tumour cell, theoretically increasing the chance of response to a mutation-targeted agent. Traditionally, germline mutations were used to identify cancer risk, however, there is growing evidence that germline variants are both prognostic and predictive of efficacy and safety treatment responses. Previous studies have determined that the presence of a germline mutation may affect overall survival independent of therapy. There is evidence of a prognostic effect for *BRCA1* and *BRCA2* mutations in ovarian cancer [37, 38] and for mismatch repair gene mutations in colorectal cancer [39-41]. In addition to their prognostic potential, germline mutations may also be predictive of response to therapy. Several trials have confirmed an increased rate of tumour response (reduction in tumour size) in patients with *BRCA* mutations [42-44]. Conversely, a mutation may predict lack of response, such as the evidence suggesting that taxane therapy might be less effective in certain *BRCA* carriers [45-47]. One major advantage of using germline variants to guide clinical care over somatic variation is that the actionable variants will remain fixed for a patient throughout their lifetime.

Now that testing is becoming cheaper and faster, genotype data for a cancer patient may be already available or is easily attainable. The challenge is to identify genetic variants that are relevant for cancer treatment, and then to integrate germline variation into routine cancer care across different oncology landscapes in order to maximise the therapeutic potential of this information [48].

1.4 Platinum therapy

The ‘platinum’ chemotherapy drugs cisplatin, carboplatin and oxaliplatin are commonly used for the treatment of lung, colorectal, ovarian, breast, head/neck, bladder and testicular cancers [25]. The first platinum-based chemotherapy drug discovered by researchers was cisplatin, which was

approved for the treatment of both ovarian and testicular cancer in 1978 [49]. In 1989 a second platinum agent, carboplatin, was approved for treatment of ovarian cancer, and due to subsequent clinical trial results the U.S. Food and Drug Administration (FDA) approved carboplatin for use treating patients with non-small cell lung cancer [50]. Finally, in 2002, a third-generation platinum drug, oxaliplatin, was approved for treatment of metastatic colorectal cancer [51]. Due to the success of platinum agents in their approved indications they are frequently used by oncologists 'off-label' for the treatment of other cancers [52]. Consequently, cisplatin is the most commonly used chemotherapy drug in the USA [53]. Yet the use of platinum therapy is not without risk, the potential antineoplastic benefit of these frequently prescribed drugs is often compromised by severe side effects including gastrointestinal upset and neutropenia. Cisplatin is regarded as the platinum drug with the most severe side effects, including nausea and vomiting, neurotoxicity, nephrotoxicity (kidney damage), and ototoxicity (hearing damage) [25, 54-56]. Beyond the immediate debilitation associated with an adverse event, patients experiencing severe symptoms often discontinue the original treatment regimen in an effort to resolve the adverse event. This discontinuation may be a dose reduction, a dose delay or complete withdrawal from platinum therapy, all of which potentially compromises a patient's benefit from the platinum compound.

To minimise the risk of adverse events while taking advantage of the antitumor activities of the platinum agents, efforts to identify safety biomarkers are under way. Two previous studies [57, 58] have identified variants within the *GSTP1* gene associated with an increased risk of neurotoxicity in patients treated with platinum therapy. In contrast, Nagashima et al. [59] identified a variant within the *SCN1A* gene that was associated with decreased risk of neurotoxicity.

Despite these efforts, we have only begun to identify genetic markers which determine the association between platinum agents and toxic side effects. In this thesis, we use data from a collection \ number of patients from Guy's and St. Thomas' NHS Foundation Trust treated with platinum therapy for lung and ovarian cancer. The patients represent a cross section of the individuals treated within a metropolitan care centre and therefore provide data from real-life routine practice conditions. This study aimed to identify association between exome-located SNPs and susceptibility to adverse drug reactions with the clinical goal of using genetic markers to guide the choice of platinum therapy [60,

61]. We obtained permission to use this data to explore the relationship between efficacy and safety by identifying the baseline characteristics and genetic factors that predict these factors.

Intensive cancer chemotherapy dosing is a complex interplay between chemotherapy complications and death from progression. For example, in acute myeloid leukaemia (AML) it is estimated that 10% of patients die from treatment related adverse events rather than progression of the underlying disease [62]. The dose intensity is recognised as a crucial factor in a patient's response to cytotoxic drugs [63]. In general, the higher the dose intensity, the greater the probability that an efficacious dose will be achieved [64]. Neutropenia is one of the most important dose-limiting toxicities of cytotoxic cancer therapy. Patients experiencing neutropenia are often given a therapy dose below their initial dosage [65] until the neutropenia has resolved. Therefore, many oncologists might consider that the absence of neutropenia a positive with respect to patient outcome as the probability of achieving an adequate tumour response will be enhanced in the absence of a dose reduction. However, since the 1990s, several studies have reported that neutropenia or leukopenia occurring during chemotherapy is positively associated with improved survival outcomes in women with breast cancer [66-69]. A meta-analysis of 1265 patients with advanced non-small-cell lung cancer (NSCLC) pooled from three randomised trials confirmed a positive association between chemotherapy-induced neutropenia and increased survival [70].

The introduction of patient genetic information has led to a further paradigm shift within oncology as researchers seek genetic markers that could predict patient response. Some researchers [19, 71] have focused on efficacy response, and propose that the risk of an adverse event is limited to only those patients who stand to derive meaningful benefit from therapy. A similar but separate incorporation of patient genetics views a drug as equally beneficial to all patients; where the more drug that can be given, the better the patient outcome. Such researchers [72, 73] have focused genetic markers of adverse events; and propose that dose be maximised in patients not predisposed to side effects thereby improving patient outcomes. The analysis in this thesis tests for genetic association with both efficacy and adverse events, and assesses evidence for genetic correlation between these outcomes.

1.5 Repeated measures studies

Repeated measures designs are a type of longitudinal trial design characterised by the collection of the trial endpoint or outcome for each patient repeatedly through time [74]. As with all longitudinal studies, the primary goal is usually to explore the change in response over time and the factors that influence change. The analysis of longitudinal data has been, and remains, a major topic in the statistical literature. The paired t-test and the analysis of variance (ANOVA) for repeated measures both represent early longitudinal analysis methods and highlight the reason for the popularity of longitudinal trial designs. Rather than model the group means at the final time point, as would be done in a simple t-test or ANOVA, the paired/repeated methods exploit the ability to measure change in response at the individual level and thereby allows patients to act as their own control. By modelling the change in response across time points, the variability in the response within groups is reduced, resulting in increased power to detect group differences in the pattern of change.

However, the benefit of a repeated measures design is not without cost and there are several statistical challenges created when collecting and correctly analysing longitudinal data. One aspect of longitudinal data that complicates the statistical analysis is that repeated measures on the same individual are usually positively correlated. This correlation violates the assumption of independence of errors. The recognition that failure to account for this correlation may result in biased results led to the formulation general linear models with correlated error structure [75] and also correlated random effects structure [75].

Clinical trials are prospective studies which often employ longitudinal measurement designs to answer questions regarding patient interventions. Within the many clinical trial designs, randomised controlled trials (RCTs) are generally considered the gold standard for evaluating the effects of new interventions or established interventions in novel disease settings. RCTs are the most robust method for establishing whether a relationship exists between treatment and outcome and for assessing the relative effectiveness of alternate treatments [76].

The fact that participants are randomised to the treatment arms is meant to ensure that, at least in expectation, the treatment groups will be balanced with respect to baseline characteristics that

influence prognosis other than the treatment being considered. Consequently, differences in outcomes between the two groups can be attributed to the effect of the differences in intervention rather than baseline characteristics.

Statistically speaking, this means that the treatment effect estimate from trial is *unbiased*, even without adjusting for any baseline covariates. Furthermore, it has been argued that this remains true even when there appears to be an imbalance between treatment arms with respect to the baseline response; and only the expectation of balance is required [77]. This is because even if imbalance is observed, it is impossible to disentangle whether there has been a failure of the randomisation process, or whether the imbalance at baseline is due to chance [78].

In reality, the 'expectation' of no imbalance is very hard to achieve. When planning a randomisation, statisticians attempt to identify the key prognostic factors related to outcome to achieve a balance between treatment arms with respect to these important factors [79, 80] through the use of a stratified randomisation. Stratified randomisation differs from simple randomisation (randomisation based only on treatment arm allocation probability) in that patients who enter a clinical trial are first grouped into strata per clinical features that may influence outcome risk. Within each stratum, patients are assigned to a treatment arm using simple randomisation schedules [81, 82]. Stratification factors are usually identified through a combination of clinical experience and prior data. If the relevant prognostic factors are unknown or poorly identified, then it is reasonable to expect that there may exist a factor imbalance between treatment arms. Even if all the relevant factors are well understood the randomisation may fail (creating treated groups that are unbalanced for prognostic features, including the baseline measurement of the outcome variable) as randomisation usually incorporates only a few of the key prognostic factors. The reason for this is that the subsequent analysis should reflect the design of the study and consequently stratification variables should be adjusted for in the analysis in order to obtain the correct type I error rates [83, 84]; and yet guidance from the European Medicines Agency (EMA) states that: "No more than a few covariates should be included in the primary analysis. Even though methods of adjustment, such as analysis of covariance, can theoretically adjust for a large number of covariates it is safer to pre-specify a simple model" [85]. Consequently, even in well understood disease models where there is clear evidence of multiple prognostic factors, less than a handful are ever incorporated into the trial as randomisation factors.

Prognostic factors not incorporated into the randomisation have the potential to create baseline imbalance between treatment arms. Other factors that may result in a failed randomisation include: 1) small sample size [86], 2) the timing of analyses (interim analyses may not have recruited all patients and failed to achieve the treatment group balance that would be manifest at full sample size), and 3) prolonged recruitment periods, which may result in changes over time with respect to prognostic characteristics of enrolled patients [81].

No matter the reason, it is fair to conclude that randomisation does not guarantee a baseline balance between treatment arms with respect to the response variable or other prognostic factors; in any individual trial, there may be large imbalances in important prognostic covariates between treatment groups merely by chance. While a pronounced baseline imbalance is not expected *a priori* in a randomised trial, when a baseline imbalance is observed *post-hoc* then researchers are faced with two options for how to proceed: 1) assume that the randomisation has worked correctly and assume that the imbalance in a random phenomenon that can be ignored, or 2) include the baseline measure as a covariate.

It is well accepted [78] that imbalance can lead to spurious estimates of the treatment effect if not accounted for in the analysis. Perhaps for this reason, EMA guidance suggests that "When the analysis is based on a continuous outcome…the baseline value should be included as a covariate in the primary analysis." [85]. Despite this regulatory guidance, in repeated measures clinical trials the decision of whether to adjust for the baseline measurement remains controversial [77, 87].

1.6 Chapter outline

This thesis covers research in two distinct fields corresponding to the areas introduced above: personalised medicine in chemotherapy, using data from a cohort of patients from Guy's and St. Thomas' NHS Foundation Trust, and an assessment of the different statistical methods used to analyse repeated measures data, with simulated data (see Figure 1.6.1).

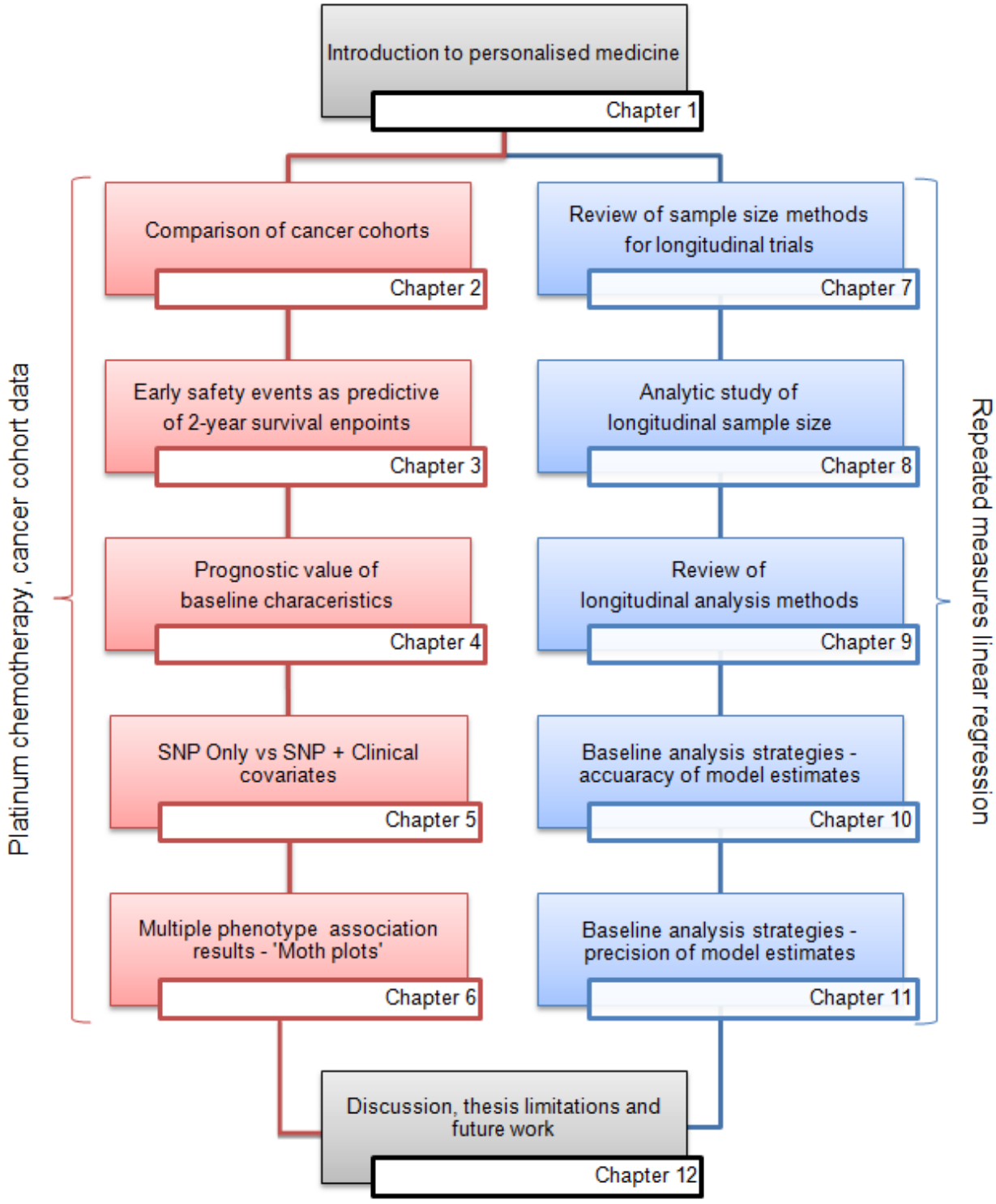


Figure 1.6.1 Schematic of thesis chapters.
Illustration of the two research lines pursued in this thesis and their relation to the chapter structure.

Chapters 2 through 6 of the thesis summarise a series of studies examining the role of baseline (demographic and disease) and genetic factors on efficacy and safety responses of Guy's and St. Thomas' NHS Foundation Trust cancer patients treated with platinum therapy.

Chapter 2 introduces the clinical data used in this section of the thesis. We describe the similarities and differences between two cohorts of data collected from patients with lung and ovarian cancer, and discuss the potential to combine the data from both cohorts through meta-analysis in subsequent association analyses.

Chapter 3 outlines our hypothesis for a link between efficacy and safety before exploring whether safety responses during treatment are predictive of landmark (2-year) overall survival and/or progression free survival. This chapter draws attention to the alternate methods and limitations of efficacy prediction from safety events.

Chapter 4 investigates the baseline factors which are predictive/prognostic of safety and efficacy responses. In this study, we tested several patient characteristics for their prognostic value in relation to both safety and efficacy outcomes to identify subgroups of patients who may be at high risk for a safety event and/or those more likely to have a positive efficacy response. We compare our results to previously identified prognostic factors and discuss the similarities and differences between our findings and previous studies in similar patient populations.

Chapter 5 explores whether SNPs, either alone or combined with baseline covariates could be used to predict patient responses. Several recent pharmacogenomic studies have identified SNPs associated with toxicity and overall survival in lung cancer patients. However, these recent genetic association studies only used the genomic data without adjusting for established clinical covariates that are known to have predictive value. We explore whether the inclusion of clinical covariates alters the top SNP-outcome associations as compared with analyses that exclude clinical covariates in searching for SNPs associated with the outcome.

Chapter 6 introduces a novel framework for exploring the similarities and differences of genetic association results from individual phenotype analyses. Each efficacy or safety response can be thought of as a phenotype. The most common (and simplest) way to deal with multiple phenotypes

is to test each SNP against each phenotype one at a time, testing for association between SNP and efficacy or safety response. We develop a graphical approach that captures both correlation and distributional differences between association results from analyses of individual phenotypes.

In Part 2 of the thesis, Chapters 7-11 explore how correlated repeated measures response data generate special challenges and opportunities for researchers in terms of the methodologies for analysis and sample size calculation.

Chapter 7 provides a review of sample size methods for repeated measures experiments. Repeated measurements collected from the same subject tend to be correlated, and the correlations must be accounted for in calculating the sample size. Failure to account for intra-patient correlation between the repeated measurement outcomes, or incorrect specification of the correlation, can result in erroneous sample size estimates. In this chapter, we review methods for the estimation of sample size in repeated measurement designs where the baseline measurement is analysed as part of the response and when the baseline measurement is used to control for individual differences through the analysis of change scores or the analysis of covariance (ANCOVA). Furthermore, we emphasise the distinction between repeated measures designs with two or with more than two time points.

Chapter 8 builds upon the material introduced in the previous chapter to explore sample size calculations for repeated measures data. Here we introduce the concept of correlation structure, a critical component of all repeated measures analysis. This study examines the influence of the strength of the correlation structure and the number of repeated measures assessments upon the required sample size to achieve a given level of power for a fixed type-I error level.

Chapter 9 reviews some of the major advances in the analysis of repeated measures data. Both marginal and mixed-models for repeated measures data model the covariance structure of the repeated measurements and represent common analysis methods for repeated measures data. In this review, we present a basic overview of each of these repeated measures analysis methods. Furthermore, we review literature which has compared alternate strategies for handling the baseline measurement within the analysis of repeated measures data.

Chapter 10 describes an extensive simulation study in repeated measures. While general linear models have emerged as the standard for analysing repeated measures data, there are several commonly employed strategies for handling the baseline measurement: 1) retain it as part of the outcome vector (RMA); 2) use the baseline measurement as a covariate in the analysis of the post-baseline measurements (ANC); 3) subtract the baseline measurement from all of the remaining post-baseline measurements and then analyse the change scores (CSA). Using simulation, we compare the model coefficients returned from each strategy to design parameters (i.e. treatment intercepts and slopes) in the simulated data.

Chapter 11 continues the comparison between RMA, ANC and CSA analysis methods. In this chapter, our focus switches to the variance of model parameter coefficients. Using matrix algebra and simulation, we explore the similarities and differences between analysis methods as they pertain to variance of the coefficients and statistical power to detect significant model effects.

Finally**, Chapter 12** summarises the contribution of these studies with a discussion of the key limitations of the work and potential future research directions.

Chapter 2. A comparison of efficacy, safety and baseline characteristics between lung and ovarian cancer cohorts

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
>
> – John W Tukey

2.1 Abstract

Background

Platinum agents form the backbone of first-line chemotherapeutic treatment regimens for both lung and ovarian cancer patients. Despite proven efficacy, platinum use carries a high risk of toxicity. Baseline characteristics in patients with cancer may predict response to therapy, including the risk of experiencing an adverse event (safety outcome / toxicity). We compared the baseline of two cohorts of cancer patients (lung and ovarian), both of which contained patients treated with platinum chemotherapy, for differences with respect to established prognostic factors for safety and efficacy outcomes.

Methods

Data from 566 patients were used to compare two cohorts of cancer patients – lung and ovarian – for similarity between baseline and demographic characteristics, safety profiles and survival characteristics. The cohorts were compared using independent t-tests for continuous characteristics. Ordinal logistic regression, generalised logit models, and chi-square tests were used to compare categorical characteristics between cohorts, as appropriate for the specific baseline characteristic.

Results

Differences between cohorts were observed for the proportion of patients in each performance scale category and with respect to the relative proportion of patients within each cohort treated with cisplatin and carboplatin containing regimens. The frequency of severe infection (p=0.0331) and fatigue (p=0.0156) was higher in the in the lung cohort as compared with the ovarian cohort. Lastly the ovarian cohort had significantly longer overall survival (p<0.0001) and progression free survival (p<0.0001) as compared with the lung cohort.

Conclusion

Despite the similarity of treatment regimens, patients from the ovarian and lung cohorts differed significantly with respect to response profiles. Future work exploring genetic associations with safety or survival outcomes cannot combine the data from both cohorts as the clinical heterogeneity between cohorts would potentially introduce confounding.

2.2 Introduction

Prognostic factors are variables that associate with subsequent clinical outcome in people with a given disease [88, 89]. Such factors include baseline patient characteristics (i.e. age, gender, ethnicity), patient behaviours (smoking status) and genetic factors. Robust identification of prognostic factors is a crucial component of personalised medicine as they are the building blocks of risk prediction models [90] and can be used to predict treatment response [89, 91]. As noted in section 1.2, there is a distinction between 'prognostic' and 'predictive' factors. Within this thesis, since all patients received similar therapy (platinum agents), we are not able to differentiate between prognostic and predictive factors, and as such the terms are used interchangeably.

Within the context of clinical trials, the similarity of comparison groups is often assessed through the baseline and demographic characteristics of comparison groups [79]. This comparison allows reviewers to make up their own mind as to whether the comparison groups were similar with respect to important prognostic factors (variables that may influence the outcome) at the start of the study. If patients are similar with respect to prognostic characteristics, then any difference in the response can be attributed to the treatment differences between comparator groups. Similarly, if patients are receiving the same chemotherapeutic agent then differences in the response can be potentially be attributed to differences in the baseline, demographic and genetic characteristics between comparator groups. In this regard, safety and efficacy outcomes are an opportunity to assess the similarity of two or more sample populations.

When a treatment is described as safe it does not mean there are no potential risks associated with its use, only that the benefits to the patient of taking the treatment are deemed to outweigh the potential risks. An adverse event (AE) is any unfavourable, unintended symptom, sign (for example an abnormal laboratory finding) or disease associated with the use of treatment. Pharmacovigilance (PV), also known as drug safety, is the pharmacological discipline related to the collection, detection, assessment, monitoring, and prevention of adverse effects with pharmaceutical products [92]. The weight of evidence regarding the risk of particular adverse events for a given therapeutic agent is known as the 'safety profile'.

The safety profile of a drug is thought to relate to both the physical and chemical properties if the drug in question as well as the process of by which the human body responds to the administration of the drug – adsorption, metabolism and excretion (ADME). This implies that the same drug will have a similar safety profile in patients with similar ADME properties. However, it has been established that demographic characteristics influence an individual's ADME. A number of age related changes in drug ADME contribute to differences in treatment response between younger and older patients. As a person ages, changes in body composition are associated with a rise in the volume of distribution for lipid soluble drugs and a reduction in the volume of distribution for hydrophilic drugs [93]. Furthermore, age related changes in renal mass might result in reduced ability to metabolise drugs and clear metabolites [94]. Beyond age, it is conceivable that alternate forms of cancer as well as cancer stage (tumour burden) might result in response differences to the same therapeutic agent.

Platinum-based agents are chemotherapeutic agents used as therapy across a range of cancers. These platinum complexes react in the body, causing DNA strands to crosslink and ultimately trigger cell apoptosis. The various platinum-based chemotherapy drugs are used against advanced, metastatic forms of colon cancer, small cell and non-small cell lung cancer, breast cancer, adrenocortical cancer, anal cancer, endometrial cancer, non-Hodgkin lymphoma, ovarian cancer, testicular cancer, melanoma as well as head and neck cancers.

Platinum based chemotherapy drugs are known to elicit strong side effects including neurotoxicity, nephrotoxicity (kidney damage), ototoxicity (hearing damage), nausea, fatigue and vomiting. Several studies have highlighted differences in the safety profile of separate platinum agents. Two large phase III trials have compared carboplatin plus paclitaxel with cisplatin-based combinations in patients with non-small cell lung cancer. Both trials demonstrated similar efficacy between carboplatin and cisplatin, but lower rates of nausea, leukopenia, and nephrotoxicity with the use of carboplatin as compared with cisplatin [95, 96]. Similarly, three trials have investigated the equivalence of carboplatin and cisplatin in combination with paclitaxel for first-line treatment of patients with ovarian cancer. All three trials concluded that carboplatin is associated with significantly lower neurotoxicity and renal toxicity [97-99].

Below we characterise the similarities and differences in baseline characteristics, safety and efficacy responses between two cohorts (lung and ovarian) of cancer patients treated with platinum containing regimens. Participants in both cohorts were recruited from the same treatment centre (oncology outpatient clinics at Guy's and St. Thomas' NHS Hospitals Foundation Trust); under the same research protocol and all patients received a platinum agent to treat their cancer (similar chemotherapy regimens). Ultimately, we wish to use the data from both cohorts to search for genetic variants associated with either safety or efficacy responses. From an analysis perspective, we would like to be able to combine the data from both cohorts thereby maximising our statistical power to detect SNP-response associations. Towards this end we need to explore the feasibility of combining the data from both cohorts. Specifically, we wish to compare cohorts for prognostic baseline characteristics, the frequency and severity of adverse events and similarity of efficacy profiles.

2.3 Materials and methods

2.3.1 Data collection

2.3.1.1 Patients and recruitment

The data sets described in this and subsequent chapters were collected as part of a study performed at Guy's and St. Thomas' NHS Foundation Trust (Principal Investigators Jeremy Sanderson, Anthony Marinaki, James Spicer) to identify genetic variants associated with platinum toxicity [100]. The study was funded by Guy's and St. Thomas' Charity and the Purine Metabolic Patient Association. The study was approved by the Regional Ethics Committee (10/H1109/47) and all participants provided written informed consent [100]. Two independent cohorts of platinum treated patients with lung and ovarian cancer were recruited. Both these cancer types receives platinum agents as first-line therapy and it was hypothesised that germline genetic variants associated with platinum toxicity would exert their influence largely irrespective of the specific cancer type [61].

Patients were recruited from oncology outpatient clinics at Guy's and St. Thomas' NHS Foundation Trust by Dr Adele Corrigan, Dr Sasala Wickramasinghe and Teagan Hoskin. Eligible patients were identified based upon planned or current treatment with a platinum containing chemotherapy regimen. Additionally, all participants considered for inclusion were ≥18 years old, exhibited adequate

haematological status to receive therapy and had histologically confirmed carcinoma. All patients were platinum-naïve prior to this study and up to four cycles of platinum-based chemotherapy were recorded. Patients within the lung cancer cohort were treated with either cisplatin (75-80 mg/m^2) or carboplatin (area under the curve (AUC) 5/6) in combination with a concomitant agent. Patients in the ovarian cohort were treated with carboplatin alone at AUC 5/6, or in combination with paclitaxel.

2.3.1.2 Clinical variables

Per institutional treatment practice, prior to each cycle of therapy, patients were evaluated for treatment tolerance. Patient outcomes and characteristics were then collected through manual review of the patient electronic records. Review of patient records is believed to be the most useful method for estimating the rate of adverse events among patients [101, 102]. Although this method is labour-intensive and cannot offer real-time monitoring of safety events, it is an appropriate method for retrospective analysis as employed in both the original work with this data and the analyses conducted in this thesis.

All adverse events were graded in accordance with the National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) Version 4.7.

2.3.1.3 Genotyping platform

At treatment initiation, consenting patients provided a blood sample from which they were genotyped. The choice of genotyping platform has been described previously [60, 100] . In brief, the choice was a function of the hypothesis that coding region variants are likely to have high pharmacogenomic relevance. The Illumina® Human Exome v1.1 DNA Analysis BeadChip Kit (Illumina UK Ltd., Saffron Walden, UK) contains >240,000 exon located and near gene polymorphisms. Content for this chip was derived from the results of 12,000 whole genome or whole exome sequences gathered from an international consortium whose research interests included depression, obesity and cancer. This chip was chosen as the platform on which to genotype all patients from the separate cancer cohorts. Genotypes were called using GenCall Data Analysis Software v1.0 clustering algorithm in Genome Studio v2011.1 (Illumina). Genotyping was conducted at the King's Genomics Centre. Both genotyping and calling was performed on this study prior to my involvement.

Several quality-control filters were applied to samples and variants prior to analysis.

Individuals were excluded for the following reasons: sex mismatch, individual call rate of >97%, SNP call rate of >99%, Hardy–Weinberg Equilibrium *p*-value $<10^{-6}$, cryptic relatedness and minor allele frequency >0.05. Population stratification was assessed using principal components (PCs). Principal components analysis was carried out using EIGENSOFT v3.0 on all common SNPs (minor allele frequency 0.05). EIGENSOFT performs principal component analysis on the $genotype \times individual$ matrix. The observed structure within the matrix usually reflects slight differences in genetic ancestry between study participants, which is captured by the eigenvector loadings for individuals. Genetic association studies standardly use the principal components (eigenvector loadings) as covariates within a regression model. These covariates model any correlation between outcome variable and ancestry, and reduces the risk of false positive results arising from this population stratification. All quality control and principal components analysis was completed by Dr Jemma Walker (Statistical Genetics Unit, King's College London).

2.3.2 Data available for analysis

The primary analysis of this data set was performed by Drs Adele Corrigan and Jemma Walker [60]. The principal investigators generously allowed me access to the data set to perform secondary analysis and for methods development. The data collection process outlined above provided me with 4 distinct datasets:

1) Lung clinical characteristics
2) Ovarian clinical characteristics
3) Lung SNP data
4) Ovarian SNP data

Below we describe the scope and format of the data within each of these 4 datasets.

2.3.2.1 Clinical data

The general structure of both clinical datasets was a ‘wide’ format (i.e. 1 row per patient). The lung cohort dataset contained 396 unique records while the ovarian cohort dataset contained 230 records.

The columns of these datasets contained baseline, efficacy and safety variables. Owing to minor differences in the safety variables collected for each cohort (see section 2.3.2.1.3), the exact number of columns differed between each dataset. We discuss each category (baseline, efficacy and safety) of variable below, and unless otherwise stated the variables mentioned are present in both lung and ovarian clinical datasets.

2.3.2.1.1 Baseline characteristics

The baseline characteristics collected were: age, gender, renal impairment graded using CTCAE, ECOG performance status, ethnicity and smoking status. Below we explain some of the baseline characteristics and methods of adverse event coding for readers less familiar with oncology data.

Performance status (PS) is an assessment of a patient's level of physical function and ability to take care of themselves [103]. The ECOG Scale of Performance status is one such measurement [104]. The ECOG performance status is an ordinal measure between 0 and 5, with 0 representing no restriction as compared with prior to the disease, and 5 representing death (see Table 2.3.1).

Table 2.3.1 ECOG performance status†

GRADE	Grade Description
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g. light house work, office work
2	Ambulatory and capable of all self-care but unable to carry out any work activities; up and about more than 50% of waking hours
3	Capable of only limited self-care; confined to bed or chair more than 50% of waking hours
4	Completely disabled; cannot carry on any self-care; totally confined to bed or chair
5	Dead

† Developed by the Eastern Cooperative Oncology Group, Robert L. Comis, MD, Group Chair.

The National Cancer Institute (NCI) of the National Institutes of Health (NIH) has published standardised definitions for adverse events (AEs), known as the Common Terminology Criteria for Adverse Events (CTCAE, also called "common toxicity criteria") [105], to describe the severity of toxicity for patients receiving cancer therapy (see Table 2.3.2). These criteria are used for the management of chemotherapy administration and dosing, and in clinical trials to provide

standardisation and consistency in the definition of treatment-related toxicity. In addition to providing consistent interpretation of severity between clinicians within a particular adverse event, the common criteria also allow for severity comparisons between adverse events. By convention, and used here, any adverse event (AA) of CTCAE grade 3 or greater is considered 'severe'.

Table 2.3.2 Grades of CTCAE

Grade 0	Asymptomatic
Grade 1	Mild symptoms, no course of action is required to treat the symptoms of the AE
Grade 2	Moderate; for this grade of AE minimal, intervention may be indicated, for example, this might involve treating the symptom (e.g. nausea and vomiting caused by a cancer treatment managed/prevented by giving the patient drugs to stop sickness, rather than stopping the anticancer drug)
Grade 3	Severe but not immediately life-threatening; for such AEs hospitalization and investigation/management are often indicated
Grade 4	Usually life-threatening consequences; for these AEs urgent intervention is indicated
Grade 5	Death related to AE

2.3.2.1.2 Efficacy

The efficacy variables collected were: date of treatment initiation, death date/date of last patient contact, progression date / date of last scan with no disease progression. Additionally, the data contained censoring markers for each of these date variables to differentiate between censors and events. Progression free survival (PFS) was calculated as the interval from the date of first dose until the date of documented disease progression or death (whichever occurred first). Patients without documented progression or death were censored at the time of the last disease evaluation. Similarly, the overall survival (OS) was calculated as the interval from the date of first dose until the date of

death from any cause. If a patient was lost to follow-up, the patient was censored on the date of last contact.

2.3.2.1.3 Safety

In contrast to the majority of clinical trials which collect all adverse events during the treatment period, the data collection process employed for these data targeted specific adverse events hypothesised to be associated with treatment. Across both data cohorts, severity information was collected for 19 specific adverse event terms (see Table 2.3.3) over 4 cycles of treatment. Consistent with the wide data format, within each dataset, specific adverse event status was represented by a column for each cycle (i.e. C1_Alopecia, C2_Alopecia for cycle 1-Alopecia and cycle-2 alopecia assessments respectively). Neither the specific date nor the treatment compliance of dosing was available. The severity of each AE at each cycle for each patient was recording using the CTCAE.

In contrast to the baseline characteristics and efficacy variables, the safety information captured from each cohort differed. Specifically, information on arthralgia, lymphocytes, myalgia and chronic kidney disease was collected from the ovarian cohort but not from the lung cohort. Similarly, information regarding dysgeusia and tinnitus was collected from the lung cohort but not from the ovarian cohort. Lastly, information regarding patient alopecia was collected across all 4 cycles in the ovarian cohort but only for the first 2 cycles in the lung cohort (Table 2.3.3).

Table 2.3.3: Adverse events collected by cycle and cohort

Adverse Event Term	***Overall***		***Ovarian Cohort***				***Lung Cohort***			
	Ovarian	***Lung***	***Cycle 1***	***Cycle 2***	***Cycle 3***	***Cycle 4***	***Cycle 1***	***Cycle 2***	***Cycle 3***	***Cycle 4***
Alopecia	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*		
Anaemia	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Arthralgia	*X*		*x*	*x*	*x*	*x*				
Constipation	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Diarrhoea	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Dysgeusia		*X*					*x*	*x*	*x*	*x*
Fatigue	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Infection	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Lymphocytes	*X*		*x*	*x*	*x*	*x*				
Mucositis	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Myalgia	*X*		*x*	*x*	*x*	*x*				
Nausea	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Neutropenia	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Peripheral Neuropathy	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Thrombocytopenia	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Tinnitus		*X*					*x*	*x*	*x*	*x*
Vomiting	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
WBC	*X*	*X*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
Chronic kidney disease	*X*		*x*	*x*	*x*	*x*				

X Information for this adverse event was collected for this cohort

Information for this adverse event was not collected for this cohort

x Information for this adverse event was collected for this cycle

Information for this adverse event was not collected for this cycle

2.3.2.2 Genotype data

The genotype data for each cohort was provided in a wide format with participants as rows and each SNP-ID forming a new column. A total of 29328 variants in 343 individuals were available for analysis from the lung cancer cohort and a total of 29028 variants and 223 individuals were available for analysis from the ovarian cancer cohort.

The number of shared variants across both cohorts was 26962.

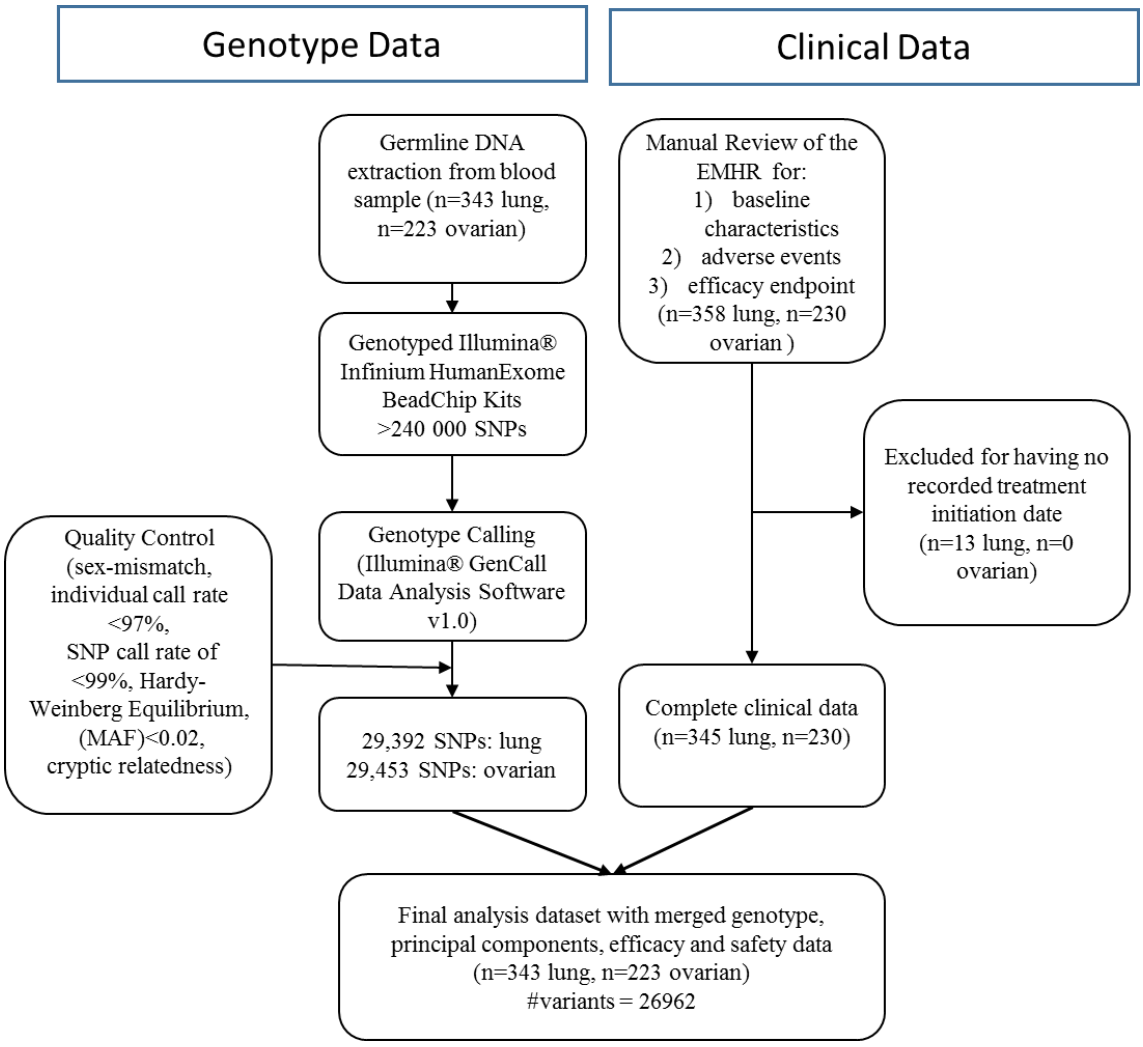


Figure 2.3.1 CONSORT diagram of patient numbers.
Schematic diagram showing clinical and genotype patient numbers by cohort. After merging the genotype and clinical data there were a total of 566 patients and 26962 SNPs available for analysis across both cancer cohorts.

2.3.3 Cohort comparison objectives

The primary objective of our study was to compare the lung and ovarian cohorts with respect to baseline patient characteristics and severe adverse events occurring at any time during the

treatment period. A severe adverse event was defined as any adverse event greater than or equal to CTCAE grade 3 that occurred following the first dose exposure and up to 30 days after the last dose of study treatment. Secondary objectives included a cohort comparison of overall survival (OS) and progression free survival (PFS).

Baseline differences in group characteristics between lung and ovarian cancer cohorts are compared using the following methods:

- An independent sample *t*-test was used to compare age between cohorts.
- The chi-square test was used to compare gender, ethnicity and smoking status.
- Ordinal logistic regression was used to analyse renal function
- A generalised logit model was used to analyse performance status. This analysis method was chosen as we rejected the null hypothesis of proportional odds when performance status was analysed as an ordinal variable.

Continuous data are described as the mean (standard deviation), median and range while the counts and percentages are presented for categorical data. Descriptive statistics were used to summarise the adverse events for each cohort and across both cohorts. Formal comparisons were made between cohorts only for adverse events occurring in greater than 5% of patients in both cohorts. These cohort comparison analyses were conducted using a chi-square statistic. Overall survival and progression free survival differences between cohorts were compared using two-sided log-rank tests. Hazard ratios were obtained from a Cox model. A 2-sided *p*-value less than 0.05 was considered significant, with no correction performed for multiple testing. All analyses were performed using SAS version 9.4.

2.4 Results

2.4.1 Baseline and demographics

Table 2.4.1 presents demographic and baseline characteristics by cohort and overall for the study population. Statistically significant differences between cohort groups were noted for age (p=0.015), performance status (p<0.0001), smoking status (p<0.0001) and renal function (p=0.020). Additionally, there were stark differences between cohorts with respect to the gender composition and platinum subclass, as expected.

Table 2.4.1 Demographic and baseline characteristics of the subjects according to cohort and overall

Baseline Characteristic	Cohort			P(a,b)
	Ovarian (N=223)	Lung (N=343)	All (N=566)	
Age(years)				0.015
N	223	343	566	
Mean(SD)	61.2 (12.9)	63.6 (10.2)	62.7 (11.4)	
Median	63.0	64.0	64.0	
Min, Max	19,88	28,88	19,88	
Gender, n(%)				N/A
Male	0	196 (57.1)	196 (34.6)	
Female	223 (100.0)	147 (42.9)	370 (65.4)	
Ethnic Origin, n(%)				0.458
Caucasian	189 (87.1)	298 (87.4)	487 (87.3)	
Asian	3 (1.4)	6 (1.8)	9 (1.6)	
Black	18 (8.3)	25 (7.3)	43 (7.7)	
Other	5 (2.3)	8 (2.3)	13 (2.3)	
Mixed	2 (0.9)	4 (1.2)	6 (1.1)	
Missing	6	2	8	
Performance status, n(%)				<.0001
0	66 (33.7)	82 (24.8)	148 (28.1)	
1	92 (46.9)	217 (65.6)	309 (58.6)	
2	29 (14.8)	30 (9.1)	59 (11.2)	
3	9 (4.6)	2 (0.6)	11 (2.1)	
4	0	0	0	
5	0	0	0	
Missing	27	12	39	
Smoking Status, n(%)				<.0001
Never	64 (57.7)	41 (12.6)	105 (24.0)	
Ex-smoker	14 (12.6)	97 (29.8)	111 (25.4)	
Current	33 (29.7)	188 (57.7)	221 (50.6)	
Missing	112	17	129	
Renal Function				
0	169 (78.6)	229 (68.4)	398 (72.4)	0.020
1	18 (8.4)	58 (17.3)	76 (13.8)	
2	26 (12.1)	46 (13.7)	72 (13.1)	
3	2 (0.9)	2 (0.6)	4 (0.7)	
Missing	8	8	16	
Treatment, n(%)				N/A
Cisplatin	0	206 (60.1)	206 (36.6)	
Carboplatin	220 (100)	137 (39.9)	357 (63.4)	
Missing	3	0	3	

Note: Percentages are based on total number of non-missing patients within each variable for each treatment group or overall as appropriate.

(a) P Value is based on Chi-Square test for the categorical variables, on 2-sample t test for the continuous variables, and using ordinal logistic regression for renal function and generalised logit model for performance status (owing to rejection of the proportional odds model when analysed as an ordinal variable).

(b) Missing values are not included in the analyses comparing cohorts for categorical variables

2.4.1.1 Age and gender

Overall mean age was 62.7 years, with the range of 19 to 88 years. In both cohorts, the majority of patients were over 60 years of age however the median and mean age were higher in the lung cohort as compared with the ovarian cohort.

2.4.1.2 Gender

In the lung cohort, most patients were male (57.1%); as expected, 100% of patients in the ovarian cohort were female.

2.4.1.3 Ethnic origin

The majority of patients in both cohorts were Caucasian (87.3% Overall – 87.1% and 87.4% for the ovarian and lung cohorts respectively). Black persons were the second largest ethnic group - 43 (7.7%). Other ethnic groups made up a very small percentage of the studied patient population.

2.4.1.4 Performance status

In general, the lung cohort had higher ECOG performance status values as compared with the ovarian cohort. The proportion of ovarian cancer cohort and lung cancer cohort patients with grade 0 ECOG performance status was 33.7% and 24.8% respectively. The number of patients with an EGOC score of 1 was 65.6% in the lung cohort and 46.9% in the ovarian cohort. The ovarian cohort had a higher proportion of patients with both PS grades 2 and 3 (14.8% and 4.6% respectively) compared with the lung cohort (9.1% and 0.6% respectively). No patients in either cohort was recorded as 4 or 5 on the ECOG scale. The cohorts differed in the number of missing performance status patients – 27 patients had missing values in the ovarian cohort as opposed to 12 patients in the lung cohort.

2.4.1.5 Smoking status

The proportion of current (57.7%) and ex-smokers (29.8%) is much higher in the lung cohort as compared with the ovarian cohort (29.7% and 12.6% respectively). The ovarian cohort had a large proportion of participants with 'Unknown' smoking status (112 - 48.1% of the cohort), which hints that the current and ex-smoker percentages may represent underestimates; however, the

ovarian cohort still has a larger proportion of never-smokers (57.7%) as compared with the lung cohort (12.6%) indicating that the proportion of current and ex-smokers is still likely to be higher in the lung cancer cohort.

2.4.1.6 Renal function

In general, the lung cohort patients had higher renal function scores (impairment) as compared with the ovarian cohort patients. The percentage of patients with a renal function score of 0 was 78.6% for the ovarian cohort and 68.4% for the lung cohort. The lung cohort had a higher proportion of patients with both renal function scores 1 and 2 (17.3% and 13.7% respectively) compared with the ovarian cohort (8.4% and 12.1% respectively). Both cohorts had 2 patients with a renal function score of 3 and again both cohorts had 8 patients missing renal function results in their records.

2.4.1.7 Treatment

In the ovarian cohort, all patients with recorded treatment received carboplatin (100%) as the platinum agent. By contrast, in the lung cohort, 60.1% of patients with recorded treatment were treated with cisplatin and 39.9% were treated with carboplatin. In both the lung and ovarian cohorts, 3 patients had missing values for their platinum chemotherapy agent.

2.4.2 Adverse events (AEs)

2.4.2.1 Brief summary of adverse events

A summary of severe adverse events by CTCAE grade is provided in Table 2.4.2. Across both cohorts, approximately half of all patients experienced at least one severe adverse event (grade ≥3): 302 (53%). The proportion of patients experiencing severe adverse events was higher in the lung cohort (196 patients - 57%) as compared with the ovarian cohort (106 patients - 48%). The proportion of patients experiencing grade 4 events in the lung cohort (16%) was approximately twice the proportion of patients experiencing grade 4 events in the ovarian cohort (8%)

Table 2.4.2 Severe adverse events by CTCAE grade

	Lung	Ovarian	Overall
		n (%)	
Any Severe AE	196 (57)	106 (48)	302 (53)
AEs of CTCAE Grade 3	183 (53)	104 (47)	287 (51)
AEs of CTCAE Grade 4	54 (16)	17 (8)	71 (13)
AEs of CTCAE Grade 5	1 (<1)	2 (<1)	3 (<1)

2.4.3 Frequency of severe adverse events

Table 2.4.3 lists the incidence of all adverse reactions of grade 3 or greater by the Medical Dictionary for Regulatory Activities (MedDRA) system order class and preferred term in each cohort and overall. Of the 566 study patients examined, 302 (53%) had at least one severe (CTCAE grade 3 or greater) adverse event during the treatment period. This group of patients had a total of 749 events indicating that some patients experienced multiple severe events, detailed further in Chapter 3. No patients experienced severe dysgeusia or alopecia.

2.4.4 Adverse events leading to death

Three patients (1 in the lung cohort and 2 in the ovarian cohort) experienced fatal adverse events (CTCAE grade 5) (see Table 2.4.2). All three fatal adverse events were recorded using the preferred term ‘infection’. The patient in the lung cohort experienced the fatal infection at cycle 2. For the two patients in the ovarian cohort the fatal infections were recorded at cycle 2 and cycle 3. These deaths contribute to the counts presented in Table 2.4.3.

Table 2.4.3 Severe adverse events by body system, preferred term, and cohort

System Organ Class / Preferred Term	Lung N= 343			Ovarian N= 223			Overall N= 566		
	Event n	Subject n(%)		Event n	Subject n(%)		Event n	Subject n(%)	
OVERALL	490	196	(57)	259	106	(48)	749	302	(53)
Blood and lymphatic system disorders	80	56	(16)	23	19	(9)	103	75	(13)
Thrombocytopenia	32	26	(8)	13	11	(5)	45	37	(7)
Anaemia	48	40	(12)	10	10	(4)	58	50	(9)
Ear and labyrinth disorders	16	9	(3)	.	.	.	16	9	(2)
Tinnitus	16	9	(3)	.	.	.	16	9	(2)
Gastrointestinal disorders	74	45	(13)	34	20	(9)	108	65	(11)
Mucositis	6	5	(1)	.	.	.	6	5	(1)
Constipation	5	5	(1)	5	5	(2)	10	10	(2)
Diarrhoea	9	9	(3)	10	7	(3)	19	16	(3)
Vomiting	25	21	(6)	11	10	(4)	36	31	(5)
Nausea	29	24	(7)	8	8	(4)	37	32	(6)
General disorders and administration site conditions	44	40	(12)	15	14	(6)	59	54	(10)
Fatigue	44	40	(12)	15	14	(6)	59	54	(10)
Infections and infestations	43	38	(11)	14	13	(6)	57	51	(9)
Infection	43	38	(11)	14	13	(6)	57	51	(9)
Investigations	232	126	(37)	152	72	(32)	384	198	(35)
Lymphocytes	.	.	.	34	14	(6)	34	14	(2)
WBC	45	36	(10)	15	11	(5)	60	47	(8)
Neutropenia	187	120	(35)	103	61	(27)	290	181	(32)
Musculoskeletal and connective tissue disorders	.	.	.	9	8	(4)	9	8	(1)
Arthralgia	.	.	.	2	2	(1)	2	2	(0)
Myalgia	.	.	.	7	7	(3)	7	7	(1)
Nervous system disorders	1	1	(0)	11	9	(4)	12	10	(2)
Peripheral sensory neuropathy	1	1	(0)	11	9	(4)	12	10	(2)
Renal and urinary disorders	.	.	.	1	1	(0)	1	1	(0)
Chronic kidney disease	.	.	.	1	1	(0)	1	1	(0)

Across both cohorts, most patients experienced at least 1 severe adverse event: 196/343 patients (57%) in the lung cohort and 106/223 patients (48%) in the ovarian cohort. The severe adverse events with the highest incidence in the two cohorts included neutropenia (Lung: 120/343 patients,

35 %; Ovarian: 61/223 patients, 27 %), fatigue (Lung: 40/343 patients, 12%; Ovarian: 14/223 patients, 6 %), and infection (Lung: 38/343 patients, 11 %; Ovarian: 13/223 patients, 6 %).

Table 2.4.4 presents the cohort comparison of severe adverse reactions reported for more than 5% of patients in both cohorts. This comparison revealed significant differences between the lung and ovarian cohorts for the incidence of fatigue (p=0.0156) and infection (p=0.0331).

Table 2.4.4 All-cause treatment-emergent adverse events occurring in ≥ 5% of patients in both cohorts

Adverse Event	**Cohort**		***p*-value†**
	Ovarian	**Lung**	
	n (%)		
Fatigue	14 (6.3)	43 (12.5)	0.0156
Infection	13 (5.8)	38 (11.1)	0.0331
Neutropenia	62 (27.8)	120 (35.0)	0.0738

†Chi-square test comparing the number of patients experiencing each adverse event type between cohorts

2.4.5 Efficacy

There were clinically meaningful differences in efficacy between the lung and ovarian cohort (see Table 2.4.5). Both the median OS and median PFS are significantly longer in the ovarian cohort as compared with the lung cohort. The OS difference between cohorts (HR=0.337) is greater than the PFS difference (HR=0.527) indicating that ovarian and lung cohorts differ not only in time to progression but also in their post-progression survival.

Table 2.4.5 Summary of efficacy outcomes

Outcome	**Cohort**		**HR† (95% CI; *p*-value)**
	Ovarian (N = 223)	**Lung (N = 343)**	
OS (days), median (95% CI)	2625 (1811-NE)	788 (685-934)	0.337 (0.257, 0.442; p<0.0001)
PFS (days), median (95% CI)	578 (498-709)	299 (267-338)	0.527 (0.430,0.645; p<0.0001)

† - HR calculated using a Cox model with the ovarian cohort as the numerator and the lung cohort as the denominator; p-value calculated using the log-rank tests stratified by cohort.
NE = Not able to estimate due to the low fraction of patients who had experienced death within the ovarian cohort

2.5 Discussion

Our results showed clinically meaningful differences between lung and ovarian cohorts with respect to prognostic baseline characteristics, the frequency of severe adverse events and the efficacy responses. The baseline data indicates two distinct differences between the cohorts 1) the proportion of patients in each performance status category and 2) the proportion of patients treated with cisplatin and carboplatin containing regimens.

Within cancer research, PS is considered a major prognostic factor [103], predicting both efficacy responses [106] as well as toxicity response to treatment [107]. Owing to this prognostic status, PS scales are commonly used within cancer research clinical trials to stratify patients at randomisation [103, 108]. Within non-small-cell lung cancer (NSCLC), the benefit derived from platinum therapy over best supportive care or single agent chemotherapy is more pronounced in 'fit' patients (PS 0 or 1) [109-111]. Consequently, differences between each cohort with respect to PS would be might result in differential survival endpoint distributions between each cohort independent of the survival prognosis associated with each disease type.

Several meta-analyses have confirmed that in the treatment of NSCLC, cisplatin and carboplatin based regimens are similarly effective [112-114]. Ardizzoni et al. [113] analysed data from 2968 NSCLC patients from nine trials conducted between 1990 and 2004. The authors observed that cisplatin-based chemotherapy was associated with more severe nausea and vomiting, and nephrotoxicity; severe thrombocytopenia was more frequent during carboplatin-based chemotherapy. Du Bois et al. [98] compared the efficacy and safety of cisplatin/paclitaxel vs. carboplatin/paclitaxel regimens in the first-line treatment of ovarian cancer patients. Similar to the NSCLC studies, the authors concluded similar efficacy between treatment arms but noted that the carboplatin containing regimen was associated with a higher frequency of hematologic toxicity, but a lower frequency of gastrointestinal and neurologic toxicity, than the cisplatin containing regimen. Within our own data, carboplatin and cisplatin use was 98.7% and 1.3% respectively for the ovarian cohort and 39.9% and 60.1% respectively for the lung cohort. This difference in the relative proportion of cisplatin use between cohorts might explain the observed cohort differences in the frequency of fatigue and nausea adverse events.

From an analysis perspective, we would like to be able to combine the data from both cohorts thereby maximising our statistical power. This benefit of pooling each cohort's individual patient data requires that cohorts: 1) contain sufficient common information for analyses, and 2) that their populations are reasonably comparable [115]. As both cohorts followed the same protocol for data collection, the first methodological concern is addressed. However, the comparability of populations is less obvious to assess. Superficially, both cohorts are comprised of patients experiencing solid tumour cancers, all of which are treated with platinum containing regimens. Additionally, the ethnic composition of each cohort is similar. However, owing to differences between cohorts in both ECOG performance status and platinum therapy subtypes, it is logical to conclude that we cannot combine cohorts in a meta-analysis of response to platinum therapy

As both cisplatin and carboplatin are thought to be equally efficacious, a meta-analysis of survival may be possible, however here the differences in the cancer types must be considered. Recent research has found that the 1-year survival rate for women with ovarian cancer following diagnosis is greater than 70%. Moreover, almost 50% of women will survive their cancer for 5 years or more [116]. By contrast for women with lung cancer, the one year survival rate is only 35.1% and the 5-year survival rate is 11.6% [116]. For men, the survival prognosis is even lower as the 1 year and 5 year survival rates are 30.4% and 8.4% respectively [116].

Recruitment to research trials is challenging, particularly for life threatening diseases. It has been estimated that fewer than 5% of cancer patients participate in clinical trials. While these data are not obtained from a prospective clinical trial each patient was required to complete a consent form and provide blood for genetic analyses and it is likely reasonable to assume that a number of patients were approached for every consent obtained. Institutional hurdles and privacy issues contribute to the labour-intensive process of gathering patient information. The data examined were part of an ambitious project to detect genetic variants associated with toxicity during chemotherapy with a platinum agent. At the time of experimental design, it was envisioned that one cohort could serve as the primary analysis and the second cohort could act as the replication cohort. Despite the similarity between cohorts with respect to the use of platinum agents as the backbone of chemotherapy treatment, we contend that differences between the cohorts with respect to the baseline characteristics, proportion of patients treated with cisplatin (rather than carboplatin) and rate of health decline (as evidenced from median PFS and OS), imposed by the

alternate cancer types, render the two cohorts' poor replication for one another. Moreover, the clinical heterogeneity of the two cohorts prohibits our ability to combine the data for meta-analysis. Consequently, in all subsequent sections of this thesis we will conduct our analyses separately in the ovarian and lung cohorts. A further limitation of the data collection protocol is the lack of cycle date for administration and/or recorded assessment of adverse event status at each cycle. This lack of date shaped the analysis that was possible. Without the specific date of each adverse event onset it was not possible to explore the time-to-onset of specific adverse event categories. Further analysis of adverse events in subsequent chapters is therefore restricted to logistic regression models.

Chapter 3. Early adverse events as predictors of mortality during platinum therapy

> Chemotherapy isn't good for you. So when you feel bad, as I am feeling now, you think, 'Well that is a good thing because it's supposed to be poison. If it's making the tumor feel this queasy, then I'm OK with it.
>
> – Christopher Hitchens

3.1 Abstract

Background

Platinum chemotherapeutic agents provide effective treatment for a variety of cancers, however most patients experience at least one adverse event (AE) during therapy. Chemotherapy induced adverse events have previously been reported to be prognostic of efficacy for some cancers although the exact relationship between non-fatal AEs and survival is not well understood. We sought to determine the impact of AEs occurring during the first 4 cycles of therapy and one-year mortality.

Methods

We examined two different adverse event types for their potential prognosis in efficacy outcomes: 1) neutropenia (NEU) and 2) gastrointestinal disorders (GID). For each adverse event patients were stratified into two groups, those experiencing grades 0-2 neutropenia and those experiencing grades 3-4 neutropenia. The efficacy outcomes studied were 2-year overall survival (OS) and 2-year progression free survival (PFS). We employed three models to explore the potential of adverse events to predict survival outcomes: 1) treating adverse event status as time-invariant; 2) treating adverse event status as time-varying and 3) the total adverse event frequency across the treatment period was used to explore if patients experiencing multiple categories of adverse events associate with poorer outcomes.

Results

No associations were found between adverse events and efficacy outcomes across any of the three analysis strategies.

Conclusions

We were unable to confirm the prognostic implications of chemotherapy induced toxicity. As the frequency and severity of adverse events are often managed prophylactically through concomitant medications, it is possible that their use in our study population confound our ability to detect an association between toxicity and efficacy. Studies assessing the prognostic potential of toxicity should consider detailed collection of concomitant medications.

3.2 Introduction

Large randomised studies have demonstrated the therapeutic benefit conferred by treatment with regimens containing either cisplatin or carboplatin for the treatment of ovarian and lung cancer [117-122]. Such studies describe benefit at the population level, however, the reality is that survival varies significantly between individual patients with some patients surviving years while others survive no more than a few days or months. In order to maximise the probability of positive patient outcomes, it is critical to have an understanding of the variety of factors which adversely affect survival, particularly beyond the first several months after termination of therapy.

Previous reports have documented the influence of baseline patient characteristics on patient mortality. Amongst the various factors identified as influencing survival in lung and ovarian cancer are: performance status [123-126], sex (lung) [123, 124], age [125, 127, 128] and lower body mass index BMI [124]. More recently there has been is shift in focus from baseline characteristics to early safety responses in the prediction of patient survival following chemotherapy [129-131]. Neutropenia is a condition characterised by abnormally low blood levels of infection-fighting neutrophils, a specific kind of white blood cell [132]. Neutropenia is a common side effect of chemotherapy [65, 133]. Most chemotherapeutic agents work by disrupting cell division [134]. While rapid cell division is a defining characteristic of cancer cells it is also a characteristic of several 'normal' cell types, including the blood cells in the bone marrow, cells in the hair follicles, and cells in the mouth and intestines [135]. Consequently, patients undergoing chemotherapy often experience chemotherapy-induced neutropenia, alopecia and gastrointestinal disorders.

Neutropenia during chemotherapy has been reported to be a predictor of better survival in patients with several types of cancers including metastatic colorectal cancer [136], adjuvant breast cancer [67], testicular cancer [137], ovarian cancer [138], non-Hodgkin's lymphomas [139], non-small-cell lung cancer [70, 140], and gastric cancer [141]. However, the relationship between neutropenia and survival is not straightforward. Neutropenia often results in a lower neutrophil-to-lymphocyte ratio, a condition that has been associated with poorer patient outcomes [142-144]. These opposing perspectives on the relationship between neutropenia and survival can best be understood by examining the dose response curve.

The dose intensity is recognised as a key aspect in a patient's response to cytotoxic chemotherapy drugs [63]. This means that both the probability of achieving an efficacy response and the probability of experiencing an adverse event both increase with dose. As safety events will limit or delay further treatment administrations (thereby depriving the patient of the therapeutic benefit conferred by the treatment), this creates a 'bell curve' relationship between dose and favourable patient response in which initial increases in dose result in improved probability of favourable patient response up to an 'optimal' dose, at which the probability of favourable patient response is maximised. Further increases in dose beyond the optimum reduce the probability of a favourable patient response (see Figure 3.2.1).

Most chemotherapy agent doses are calculated using body surface area (BSA). As volume increases as a cubic function and BSA varies with patient size as a squared function, BSA standard dosages may be too small for optimal efficacy in larger patients, or too large to avoid unnecessary adverse effects in smaller patients.

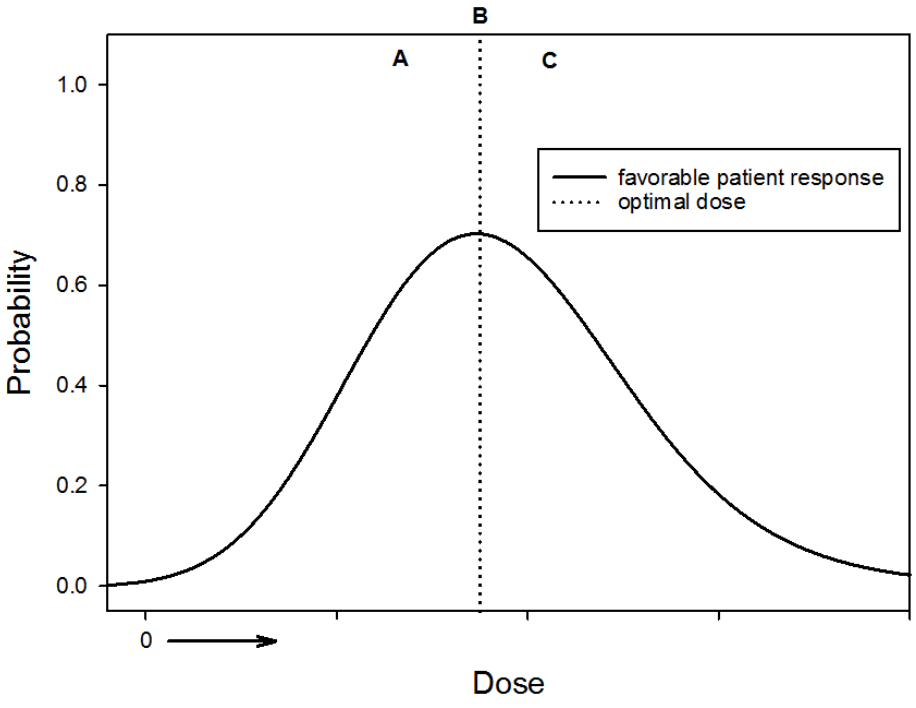


Figure 3.2.1 Dose response curve.
In section A, increasing the dose improves the probability of a favourable patient response. The point B on the dose scale represents the 'optimal' dose at which the probability of a favourable patient response is maximised. Further increases in dose (region C) decrease the probability of a favourable patient response due to AEs.

This dose-response relationship can best be understood by partitioning the efficacy and safety dose response relationships (see Figure 3.2.2).

At low dose, the probability of either an efficacy or adverse response is low. The probability of achieving an efficacy response increases with dose until sufficiency is reached, at which point further increases in dose will not improve the chances of an efficacy response. The probability of a dose-limiting adverse response also increases with dose, but at a slower rate than the efficacy curve. The difference in the slope of probability between efficacy and safety responses creates a dose window in which the probability of achieving a positive efficacy response is greater than the probability of an unwanted safety response. The difference between the dose that is efficacious and the dose that causes adverse effects is known as the 'therapeutic window'. Subtraction of the safety response curve from the efficacy response curve creates a distribution that we can recognise as the 'favourable patient response' curve.

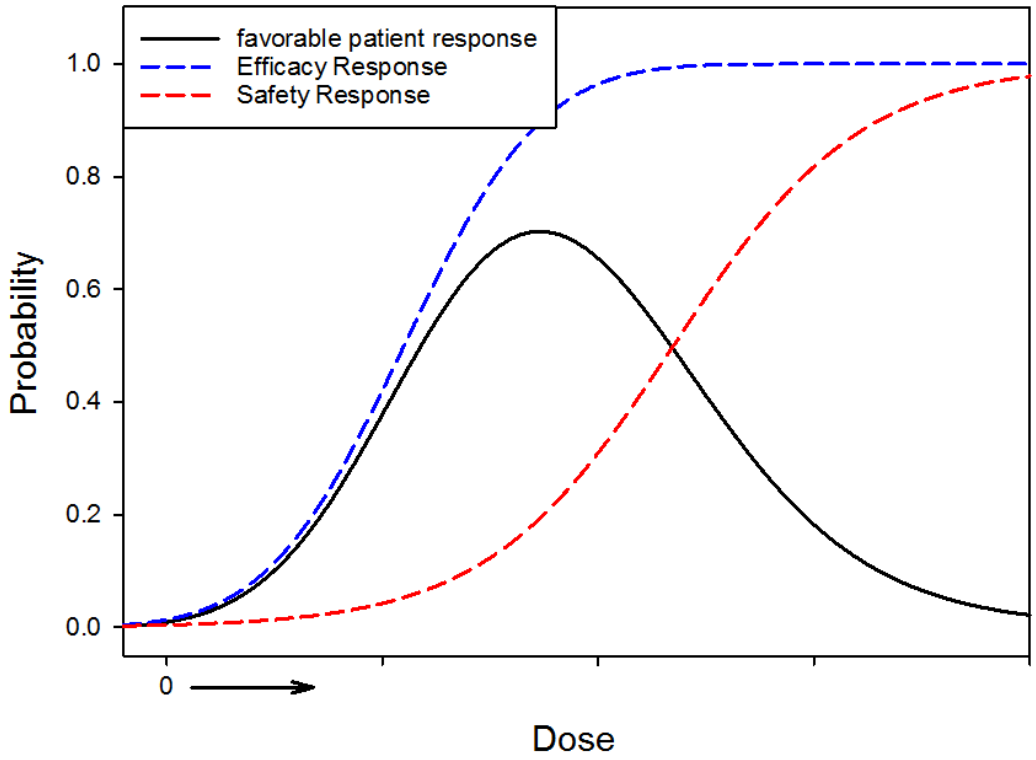


Figure 3.2.2 Dose response curves for efficacy, safety and favourable patient response.
The favourable patient response curve is derived from the subtraction of the safety response curve from the efficacy response curve. The shape of the favourable response curve is a product of the difference in slopes between the efficacy and safety dose-response curves.

These curves show that the proportion of patients who are receiving less than the optimal dose, and the proportion of patients who are receiving beyond the optimal dose, will influence whether a study finds a positive or negative association between an adverse event and survival. Irrespective of the direction of association, this framework establishes that a link between efficacy and safety endpoints is a logical assertion.

We hypothesised that clinically significant, but nonfatal, AEs that occur during the first four treatment cycles would be predictive of later-term survival.

Therefore, utilising standardised CTCAE definitions, we conducted a retrospective analysis of patients receiving platinum therapy for the treatment of solid tumour cancers to determine the separate and combined effects of early adverse events on two-year overall survival and progression free survival.

3.3 Methods

3.3.1 Adverse events

Neutropenia (NEU) and gastrointestinal disorder (GID) adverse events reported during the first four cycles of platinum therapy were included in the analyses. Adverse event information was obtained from patient responses and laboratory assessments during clinic visits. The severity of each adverse event was assessed using the CTCAE. The system organ class (SOC) was determined using MedDRA. The specific terms included in gastrointestinal disorder (GID) were: mucositis, constipation, diarrhoea, vomiting, and nausea.

3.3.2 Endpoints

Although adverse events during the treatment period may influence OS/PFS at any time beyond the treatment phase, second and third line therapy has the potential to introduce time-dependent confounding as patient OS may well be influenced by the use of post-progression therapy [145, 146]. We chose to assess the impact of toxicity responses upon efficacy at a given time point: 2 years [145]. This time-point captured the majority of OS events in both cohorts while reducing the risk for subsequent therapy to influence patient OS as compared with analysis of the raw survival times.

In this method, patients who had not experienced the survival event by the landmark time were censored in the analysis.

PFS was therefore defined as the time from treatment initiation to the earliest documentation of disease relapse or death (all causes) or end of 2-year follow-up, whichever came first. Similarly, OS was defined as the time on treatment to the date of death or end of 2-year follow-up, whichever came first.

The relationship between the occurrence of specific adverse events (NEU and GID) and efficacy outcomes (OS and PFS) was assessed. Furthermore, we evaluated the relationship between outcome measures and the number of different kinds of specific adverse events reported per patient.

3.3.3 Survival analysis

In oncology studies, the primary efficacy outcome under assessment is the time to an event of interest, such as the time from cancer diagnosis or treatment initiation to cancer recurrence or death. The event of interest may not have occurred for every patient at the time of the statistical analysis, and similarly, a subject may be lost to follow-up before the event is observed. In such cases, the data are censored at the time of the analysis or at the time the patient was lost to follow-up. Censored data still contribute information to the analysis; whilst we do not know the exact date of the event, we know that it had not occurred up until the censoring time.

The Kaplan-Meier method is a common statistical method used to analyse survival data [147, 148], and to estimate the survival probability - S(t), the probability that a subject survives beyond some time t.

The survival probability is estimated by the survival function $S(t) = P(T > t)$, where T is the survival time. The Kaplan-Meier method assumes no specific underlying function and therefore estimates the survival probability non-parametrically [148].

The survival curves can be compared statistically using either the log-rank test or a Cox proportional hazard model. While simpler to calculate, the log-rank test does not allow for the inclusion of covariates (or continuous predictors) [149]. By contrast the Cox proportional hazard (PH) model accounts for multiple categorical or continuous risk factors simultaneously [150] .

3.3.3.1 Static/constant predictors

A static predictor is any factor that does not change across time (i.e. biological sex). A Cox model incorporating only static explanatory variables can be expressed in the following way:

$$h_i(t, X) = \lambda_0(t)e^{\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}} \quad (3.3.1)$$

Where baseline hazard rate $\lambda_0(t)$ is an unspecified non-negative function of time that corresponds to the hazard rate when all covariate values are equal to zero (i.e. $X_{i1} = 0, X_{i2} = 0, \ldots X_{ik} = 0$); and $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the regression function.

If we are interested in a single categorical covariate with two levels ($X_1^*$and X_1 where $X_1^* = X_1 + 1$) then the hazard is:

$$h(t, X) = \lambda_0(t)e^{\beta X} \quad (3.3.2)$$

The hazards for patients with covariate values X_1 and X_1^* are given respectively by $h(t, X_1) = \lambda_0(t)e^{\beta X_1}$ and $h(t, X_1^*) = \lambda_0(t)e^{\beta X_1^*}$, and the hazard ratio (HR) is calculated by:

$$HR = \frac{h(t, X_1^*)}{h(t, X_1)} = e^{\beta(X_1^* - X_1)} \quad (3.3.3)$$

An important feature of equation (3.3.3) is that $\lambda_0(t)$ cancels out of the numerator and denominator and therefore the HR corresponds to the change in hazard associated with one-unit increase in the explanatory variable X on the risk of event. If the HR is greater than one (β > 1), the event risk is increased for subjects with covariate value X_1^* as compared to subjects with covariate value X_1, while a HR lower than one (β < 1) indicates a decreased risk. Although the hazard rate $h(t, X)$ will vary over time, the HR is constant; this is the assumption of proportional hazards [150].

3.3.3.2 Time-varying predictors

The model for the hazard, given in (3.3.1), involves the proportional-hazards assumption of constant/static covariate status across the observation period. This means that for each patient,

their specific covariate value will not change from baseline. While many patient characteristics should remain constant throughout the observational period of a study, others could vary across time: time-dependent (or time varying) variables. Performance status, smoking status and adverse event status are all examples of variables that could vary over time. While typically researchers hold the values of such variables fixed at a certain point in time (traditionally baseline), it has been demonstrated that allowing the predictor values to vary with time can yield a more accurate estimate of the predictor effect [151].

To extend (3.3.2) to allow the predictor X to change over time we need to create an interaction term between X and a function of time t for example ($f(t) = t, t^2, log(t)$). For simplicity, we will assume that $f(t) = t$:

$$h(t, X) = \lambda_0(t) e^{\beta X(t)} \tag{3.3.4}$$

And the HR is given by:

$$HR(t) = \frac{h(t, X_1^*)}{h(t, X_1)} = e^{\beta(X_1^*(t) - X_1(t))} \tag{3.3.5}$$

In contrast to (3.3.3), in models with time-varying covariates the HR corresponds a unit increase in the variable X but the HR is now time-dependent through the function $f(t)$. Even though the HR varies with time in (3.3.5), the coefficient (β) of the difference in values of the time dependent variable X_1 is itself not time dependent. The coefficient therefore represents an aggregate effect of the time-varying variable across the observation period [152].

3.3.4 Statistical analysis

The primary objective of this study was to evaluate the association between severe (CTCAE grade ≥3) gastrointestinal / neutropenic adverse events during four cycles of platinum therapy and the efficacy endpoints overall survival (OS), and progression free survival (PFS).

In order to evaluate the prognostic implications of each adverse event we first identified the worst grade of each adverse event during treatment for each patient. Owing to the size of the study population, we pooled each adverse event into two categories: unaffected (grades 0-2) and affected (grades 3-4). Overall survival was defined as the interval between the date of treatment initiation and the date of death or last follow-up. Similarly, progression free survival was defined

as the interval between the date of treatment initiation and the date first documented progression or death or last follow-up, whichever occurred first.

To evaluate the impact of each AE on efficacy we used 3 analyses strategies:

(1) Univariate regression analyses without time-varying covariate (TVC) - i.e. time-invariant analyses, examined the 2-year OS and PFS of patients in each cohort against adverse event status (affected / unaffected) by specific adverse event, cycle and cohort.

(2) Number of specific adverse events reported - to explore if total adverse event count during treatment period influenced efficacy, we examined the 2-year OS and PFS of patients in each cohort against total sum of neutropenic and gastrointestinal adverse events during the treatment period. Univariate analysis explored the association between efficacy endpoints and the frequency of adverse events in the absence of covariates. Multivariable analyses repeated the univariate analysis including six baseline characteristics as covariates to the adverse event count: gender, age, performance status, renal function, race and smoking status.

(3) Univariate and multivariable regression analyses with TVC – As adverse event status can vary by cycle, our first analysis could be biased by early event lead time, therefore we analysed each adverse event as a TVC where the length of each cycle was assumed to be 28 days. As with the fixed effect analyses the covariates in the multiple variable regression were: gender, age, performance status, renal function, race and smoking status.

All analysis strategies utilised the Cox proportional hazards model and therefore the measure of association in this study was the HR along with the 95% confidence interval (95%CI). In the multivariable analyses, only the main effects were entered into the model; we did not explore interaction terms between covariates (possible synergistic effects). Within the context of the software employed for the analysis (SAS v.9.4), Type III tests refer to a test of hypothesis that all coefficients in the model are 0, that is, an overall test of whether there are any differences in event rate across any of the levels of the covariate [153, 154].

As described in the previous chapter, all analyses were performed separately in the ovarian and the lung cohorts. All tests we performed with an alpha of 0.05, with no correction for multiple testing, the limitations of which are highlighted in the discussion.

3.4 Results

3.4.1 Patient characteristics - NEU, GID

Baseline characteristics of the unaffected and affected patients for each adverse event are presented in Table 3.4.1. In both cohorts (ovarian and lung) more patients experience severe

neutropenia (NEU) as compared with severe gastrointestinal adverse events (GID). Affected patients in both adverse event types had marginally higher mean age, performance status and renal function scores as compared with unaffected patients.

Table 3.4.1 Baseline characteristics for unaffected and affected patients by adverse event type and cohort

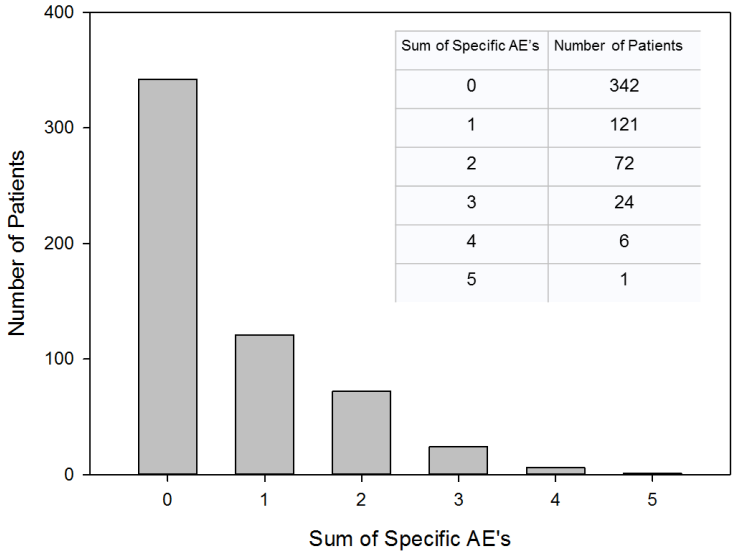
	Ovarian				Lung			
	GID		NEU		GID		NEU	
	Unaffected (N=203)	Affected (N=20)	Unaffected (N=162)	Affected (N=61)	Unaffected (N=298)	Affected (N=45)	Unaffected (N=223)	Affected (N=120)
Gender, n(%)								
Male	0	0	0	0	179 (60.1)	17 (37.8)	129 (57.8)	67 (55.8)
Female	203 (100.0)	20 (100.0)	162 (100.0)	61 (100.0)	119 (39.9)	28 (62.2)	94 (42.2)	53 (44.2)
Age								
N	203	20	162	61	298	45	223	120
Mean	61.0	64.0	60.7	62.6	63.6	63.8	62.6	65.4
SD	13.18	9.56	13.26	11.91	10.08	11.23	10.64	9.15
Max	88	78	88	86	88	79	88	84
Min	19	46	19	42	28	36	28	39
Ethnicity, n(%)								
Caucasian	172 (84.7)	17 (85.0)	139 (85.8)	50 (82.0)	258 (86.6)	40 (88.9)	191 (85.7)	107 (89.2)
Asian	2 (1.0)	1 (5.0)	2 (1.2)	1 (1.6)	6 (2.0)	0	2 (0.9)	4 (3.3)
Black	17 (8.4)	1 (5.0)	12 (7.4)	6 (9.8)	21 (7.0)	4 (8.9)	20 (9.0)	5 (4.2)
Other	4 (2.0)	1 (5.0)	4 (2.5)	1 (1.6)	8 (2.7)	0	6 (2.7)	2 (1.7)
Mixed	2 (1.0)	0	1 (0.6)	1 (1.6)	3 (1.0)	1 (2.2)	2 (0.9)	2 (1.7)
Missing	6 (3.0)	0	4 (2.5)	2 (3.3)	2 (0.7)	0	2 (0.9)	0
Performance status, n(%)								
0	63 (31.0)	3 (15.0)	49 (30.2)	17 (27.9)	73 (24.5)	9 (20.0)	50 (22.4)	32 (26.7)
1	84 (41.4)	8 (40.0)	61 (37.7)	31 (50.8)	185 (62.1)	32 (71.1)	150 (67.3)	67 (55.8)
2	23 (11.3)	6 (30.0)	22 (13.6)	7 (11.5)	26 (8.7)	4 (8.9)	18 (8.1)	12 (10.0)
3	8 (3.9)	1 (5.0)	8 (4.9)	1 (1.6)	2 (0.7)	0	0	2 (1.7)
4	0	0	0	0	0	0	0	0
Missing	25 (12.3)	2 (10.0)	22 (13.6)	5 (8.2)	12 (4.0)	0	5 (2.2)	7 (5.8)
Smoking status, n(%)								
Never	58 (28.6)	6 (30.0)	45 (27.8)	19 (31.1)	32 (10.7)	9 (20.0)	25 (11.2)	16 (13.3)
Ex-smoker	12 (5.9)	2 (10.0)	11 (6.8)	3 (4.9)	86 (28.9)	11 (24.4)	67 (30.0)	30 (25.0)
Current	31 (15.3)	2 (10.0)	25 (15.4)	8 (13.1)	166 (55.7)	22 (48.9)	120 (53.8)	68 (56.7)
Unknown	102 (50.2)	10 (50.0)	81 (50.0)	31 (50.8)	14 (4.7)	3 (6.7)	11 (4.9)	6 (5.0)
Renal Function, n(%)								
0	156 (79.6)	13 (68.4)	124 (79.5)	45 (76.3)	198 (68.0)	31 (70.5)	157 (72.4)	72 (61.0)
1	18 (9.2)	0	15 (9.6)	3 (5.1)	54 (18.6)	4 (9.1)	34 (15.7)	24 (20.3)
2	21 (10.7)	5 (26.3)	16 (10.3)	10 (16.9)	37 (12.7)	9 (20.5)	25 (11.5)	21 (17.8)
3	1 (0.5)	1 (5.3)	1 (0.6)	1 (1.7)	2 (0.7)	0	1 (0.5)	1 (0.8)

3.4.2 Association of NEU and GID with OS and PFS, by cohort and cycle

Table 3.4.2 shows the results of univariate analyses of 2-year OS and PFS against severe adverse event, by adverse event type, cycle and cohort. Each HR measures the hazard of affected over unaffected patients. Univariate Cox analyses suggested that patients experiencing severe gastrointestinal events during cycle 1 of therapy were associated with higher risk of death (*p=0.0234*; Table 3.4.2) in the lung cohort only. The HR for affected-GID patients relative to unaffected-GID patients was 1.908 (95% CI: 1.081-3.368). We did not observe any associations between specific adverse events and overall survival in the ovarian cohort. In both cohorts, cycle 1 severe gastrointestinal adverse events were associated with a shorter progression free survival time (*p*=0.0183 and *p*=0.0286 for the ovarian and lung cohorts respectively). The HR for affected-GID patients relative to unaffected-GID patients was similar between cohorts: 2.087 (95% CI: 1.117-3.898) in the ovarian cohort and 1.723 (95% CI: 1.052-2.821) in the lung cohort.

3.4.3 Number of specific adverse event association with OS and PFS

A patient could have a maximum of two types of specific AEs, which was the sum of neutropenic and gastrointestinal events (affected/unaffected) across 4 cycles. This gives a total range of between 0 and 8. The observed range was between 0 and 5. As only one patient experienced 5 specific AEs during the treatment period, this patient was counted as having 4 specific AEs in the analyses. Figure 3.4.1 shows a distribution of patient scores within each category. Table 3.4.3 presents the results of both univariate and multivariable regression analyses. The highlighted cells provide the Type III test of association results and the non-highlighted cells provide association values for each event count level of affected status patients relative to unaffected patients (event count=0). We found no evidence of association between the sum of specific AEs and efficacy outcomes in either cohort (See Table 3.4.3).



Sum of Specific AE's	Number of Patients
0	342
1	121
2	72
3	24
4	6
5	1

Figure 3.4.1 Frequency of patients in each sum of specific AEs category (i.e. sum of all neutropenic and gastrointestinal adverse events across all cycles) – pooled across both cohorts.

3.4.4 TVC adverse event association with OS and PFS

Table 3.4.4 presents the results of the univariate and multiple regression analyses when considering each severe adverse event as a TVC. In these analyses, severe adverse events were not identified as being prognostic factors for OS or PFS.

Table 3.4.2 Landmark analysis of two-year survival outcomes predicted by neutropenic and gastrointestinal adverse events

	Ovarian						Lung					
	Neutropenia			Gastrointestinal Disorder			Neutropenia			Gastrointestinal Disorder		
Cycle	**HR†**	**95% CI**	***p*-value**	**HR†**	**95% CI**	***p*-value**	**HR**	**95% CI**	***p*-value**	**HR†**	**95% CI**	***p*-value**
Overall Survival												
1	1.275	0.392-4.149	0.68596	1.907	0.677-5.375	0.21420	1.130	0.732-1.746	0.58070	1.908	1.081-3.368	0.02340
2	1.576	0.659-3.770	0.30273	N/A	N/A	0.48230	0.657	0.411-1.050	0.07701	0.868	0.384-1.964	0.73403
3	1.703	0.806-3.599	0.15795	3.059	0.417-22.43	0.24676	0.794	0.510-1.236	0.30574	0.613	0.227-1.655	0.32908
4	1.263	0.528-3.021	0.59968	3.462	0.832-14.40	0.06890	0.986	0.569-1.707	0.95943	2.478	0.789-7.785	0.10774
Any	1.785	0.931-3.422	0.07697	1.635	0.638-4.188	0.30089	0.926	0.667-1.285	0.64347	1.209	0.770-1.897	0.40922
Progression Free Survival												
1	1.561	0.815-2.989	0.17503	2.087	1.117-3.898	0.01829	0.923	0.645-1.319	0.65840	1.723	1.052-2.821	0.02864
2	1.479	0.883-2.476	0.13409	0.539	0.075-3.860	0.53170	0.800	0.568-1.126	0.19849	0.980	0.535-1.796	0.94852
3	1.566	0.998-2.457	0.04901	0.814	0.114-5.834	0.83724	0.702	0.489-1.009	0.05400	0.809	0.400-1.636	0.55319
4	1.121	0.669-1.876	0.66444	0.843	0.208-3.411	0.80979	0.943	0.614-1.448	0.78744	1.107	0.354-3.458	0.86122
Any	1.431	0.970-2.111	0.06956	1.409	0.790-2.512	0.24278	0.820	0.631-1.067	0.13817	1.167	0.809-1.683	0.40912

†HR calculated using analysis strategy (1) – Cox regression without time-varying covariate

Table 3.4.3 Two-year survival outcomes predicted by the number of specific adverse events reported

	Ovarian						Lung					
Event Count	**Univariate HR**	**95% CI**	***p*-value‡**	**Multivariable HR**	**95% CI**	***p*-value ‡**	**Univariate HR**	**95% CI**	***p*-value ‡**	**Multivariable HR***	**95% CI**	***p*-value ‡**
Overall Survival												
0	ref	-	0.348	ref	-	0.293	ref	-	0.998	ref	-	0.859
1	2.115	1.006-4.445	0.048	1.548	0.732-3.275	0.253	0.996	0.676-1.467	0.985	1.043	0.699-1.556	0.838
2	1.681	0.671-4.211	0.267	2.123	0.827-5.449	0.117	0.920	0.572-1.480	0.731	0.781	0.481-1.267	0.316
3	2.256	0.302-16.870	0.427	5.534	0.704-43.520	0.103	1.004	0.506-1.993	0.990	0.864	0.412-1.812	0.699
4	1.437	0.192-10.740	0.724	1.709	0.226-12.930	0.603	0.000	0.000-NE	0.972	0.000	0.000-NE	0.973
Progression Free Survival												
0	ref	-	0.148	ref	-	0.106	ref	-	0.749	ref	-	0.374
1	1.590	0.985-2.567	0.057	1.289	0.793-2.095	0.305	0.958	0.705-1.302	0.782	0.905	0.656-1.249	0.544
2	1.776	1.049-3.008	0.032	1.988	1.159-3.411	0.012	0.764	0.515-1.134	0.181	0.673	0.450-1.007	0.054
3	1.479	0.361-6.056	0.586	2.401	0.576-10.010	0.229	0.936	0.530-1.653	0.818	0.836	0.457-1.528	0.559
4	1.511	0.549-4.160	0.424	1.624	0.587-4.495	0.350	0.738	0.183-2.984	0.670	0.590	0.143-2.425	0.464

‡ *p*-values calculated using analysis strategy (2) – Cox regression without time-varying covariate. When the event count=0, the *p*-value is derived from the Type III test of association results; when event count > 0 the *p*-value represents association values for each event count level of affected status patients relative to unaffected patients (event count=0).
NE = Not able to estimate due to the low fraction of patients who experienced 4 severe events within the lung cohort.

Table 3.4.4 Univariate and multivariable analyses with each adverse event modelled as a TVC

	Ovarian						**Lung**					
	Neutropenia			**Gastrointestinal Disorder**			**Neutropenia**			**Gastrointestinal Disorder**		
Cycle	**HR[T]**	**95% CI**	***p*-value**	**HR[T]**	**95% CI**	***p*-value**	**HR[T]**	**95% CI**	***p*-value**	**HR[T]**	**95% CI**	***p*-value**
Overall Survival												
Univariate	1.366	0.568-3.287	0.4856	NA	NA	0.9895	0.829	0.512-1.343	0.4465	1.087	0.480-2.463	0.8419
Multivariable	1.940	0.774-4.865	0.1576	NA	NA	0.9895	0.643	0.389-1.060	0.0836	1.183	0.518 - 2.701	0.6894
Progression Free Survival												
Univariate	1.131	0.754-1.696	0.5529	0.528	0.074-3.786	0.5255	1.067	0.791-1.440	0.6721	1.645	0.896-3.022	0.1085
Multivariable	1.289	0.823 2.018	0.2674	0.518	0.071 3.797	0.5172	1.012	0.737 1.388	0.9427	1.392	0.734-2.640	0.3107

[T] HR calculated using analysis strategy (3) – Cox regression with time-varying covariate

3.5 Discussion

The objective of this study was to investigate whether chemotherapy-induced severe neutropenia or gastrointestinal disorders were related to efficacy endpoints and could therefore be used as a predictor of patient prognosis. This analysis approach has been used to identify associations prognostic between safety endpoints in many studies [67, 69, 155] and is considered to be an appropriate analysis method in treatment scenarios where the number of deaths during the treatment period is low.

The underlying mechanism of severe adverse events acting as prognostic factors could be explained by adverse events acting as a measure of the plasma drug concentration, i.e. the lack of adverse events might indicate that the chemotherapy dosage is too low to produce a toxicity response.

The present analyses showed no evidence that patients who reported one or more adverse events during the first 4 cycles of therapy had alternate subsequent efficacy outcomes in terms of 2-year landmark analysis of PFS and OS patients as compared with patients who do not report these adverse event symptoms. Only the non-time varying analysis of severe adverse events during therapy found evidence of association with efficacy at the pre-specified alpha level. The strongest observed association was between cycle 1 GID and 2-year PFS. A modest association was also observed between cycle 1 GID and worse survival outcomes in the lung cohort. There was no evidence of associations between adverse events and OS in the ovarian cohort. However, it must be pointed out that given the number of tests performed, these observed associations are likely spurious. Based upon the number of hypothesis tests performed in exploring the association between two-year survival outcomes predicted by adverse events, a Bonferroni adjusted *p*-value threshold for the of 0.05/40 = 0.00125 is a more reasonable benchmark of association. Using this critical threshold, none of the observed associations would pass correction for multiple testing (can be considered significant).

A recognised limitation of using adverse events as a prognostic measure of efficacy is that patients who survive longer tend to receive more chemotherapy, and thus have a greater chance of developing adverse events. Consequently, considering severe adverse events as a baseline factor can produce false-positive associations between safety and efficacy endpoints. Our

specific data presented us with data only on the first four cycles of chemotherapy, and while we are not specifically aware of extended treatment, some patients may have continued to receive chemotherapy beyond four cycles. In an effort to avoid spurious associations between safety and efficacy, we analysed each adverse event as a TVC [70, 136, 156]. In the analyses that considered severe adverse events as TVCs, neither neutropenia nor gastrointestinal disorders predicted overall survival or progression free survival. These findings differ from previous studies which have identified chemotherapy induced neutropenia as a being associated with improved survival in metastatic colorectal cancer [157], breast cancer [158], non-small-cell lung cancer [70, 159], gastric cancer [160], and ovarian cancer [159].

This study has several important limitations that potentially hamper our ability to detect survival and toxicity links. This investigation included patient data from medical records completed from within the same Guy's and St. Thomas' NHS Foundation Trust clinic but by alternate physicians. Heterogeneity in the reporting of adverse events is a potential source of confounding. Historically, ovarian cancer patient survival is longer than that expected for patients with lung cancer. Consequently, it is possible that clinicians pay greater attention to the reporting and management of AEs within ovarian cancer as compared with lung cancer. Because of the post-hoc nature of these analyses, concomitant medications and dose intensity were not recorded. Granulocyte colony stimulating factor (G-CSF) and granulocyte-macrophage colony stimulating factor (GM-CSF) are glycosylated polypeptides that induce an increase in the proliferation and maturation of white blood cells including neutrophils and monocytes-macrophages. Colony-stimulating factors have been shown to reduce the duration and severity of neutropenia and the risk of febrile neutropenia [161] and enable delivery of more intensive or dose-dense chemotherapy when indicated [162]. The American Society of Clinical Oncology (ASCO) release guidelines that detail how to use chemotherapy, and also concomitant medications to improve patient outcomes and quality of life through the management of adverse events. The current guidelines call for prophylactic use of CSFs to reduce the risk of neutropenia when the risk of neutropenia is approximately 20% or higher [162]. Specifically, primary prophylaxis with a CSF should start with the first cycle and continue throughout subsequent cycles of chemotherapy. Secondary prophylaxis with a CSF is recommended when a patient experiences a neutropenic event related to a prior dose of chemotherapy, in which primary prophylaxis was not given, in which the adverse event may result in a reduced dose or delay of treatment dose which may compromise disease-

free or overall survival outcomes [162]. In addition to chemotherapy regimen and type of malignancy, advanced age, poor performance status and poor renal function are all risk factors for neutropenia indicated by ASCO [163-165]. While there are practical differences between America and Britain in the care of cancer patients, it is probable the NHS recognise similar risk factors for prophylaxes. Consequently, despite the high rate of neutropenia observed in the two cohorts, CSF use might be confounding an association between severe neutropenia and efficacy outcomes.

In addition to CSF use, the frequency of adverse events is related to treatment compliance. Patients who did not experience specific adverse events may not have been receiving the full dose of the prescribed regimen owing to other medical complications.

No association was found between the cumulative incidence of gastrointestinal and neutropenic AEs upon survival. Although it is conceivable that older patients may be less likely than younger patients to report AEs, that the platinum subtype may alter the risk of AE, or even that there may be gender differences in probability and/or reporting of adverse events, after adjusting for these possible confounders, no association between specific AEs and survival outcomes was observed.

Although our results failed to show an association between specific adverse effects and efficacy outcomes, we believe that the approach has the potential to unlock meaningful associations and thus provide a valuable predictor and biomarker of treatment efficacy. Future prospective studies might seek to use a different landmark analysis time point (i.e. 3 or 5-year survival) and also use different AE categories when examining the potential for association.

Chapter 4. Prognostic factors in patients treated with platinum containing regimens for cancer

You can believe the diagnosis, not the prognosis

– Deepak Chopra

4.1 Abstract

Background

Cancer patient prognosis has two aspects: safety and efficacy. Safety prognosis describes the risk that a patient will experience an adverse event in response to treatment and efficacy prognosis describes the probability of achieving favourable survival characteristics in response to treatment.
Chemotherapy-related hospitalizations in patients with advanced cancer are both common and costly. Methods to identify patients at high risk of chemotherapy toxic effects have the potential to aid in the development of targeted treatment strategies to prevent chemotherapy-related hospitalizations and allow for forecasting of hospitalisation rates.
Patients with cancer commonly want to know the survival statistics for people in similar situations in an attempt to know what to expect. The survival estimate is based upon person's characteristics as they pertain to factors known to affect survival rates.
The current study was undertaken to explore whether baseline characteristics predicts either adverse event rates or survival in cancer patients treated with platinum containing regimens.

Methods

Data were analysed from 566 patients, 223 with ovarian cancer and 343 lung cancer, who were treated with platinum chemotherapy. Primary endpoints were progression free survival (PFS) overall survival (OS) and any severe adverse event (AA). Potential prognostic variables were included in Cox proportional hazard regression models. Multivariable analysis was conducted to identify independent prognostic factors.

Results

Baseline performance status (PS) was a significant independent predictor for both PFS and OS in both lung and ovarian cohorts. Baseline PS was also predictive of AA in the ovarian cohort while age and sex were predictive of AA in the lung cohort. Sex was also an additional independent predictor for OS in the lung cohort and renal function was an additional predictor of OS in the ovarian cohort. Smoking status at baseline predicted PFS in the lung cohort only.

Conclusion

Performance status at baseline isprognostic in both ovarian and lung cancer for efficacy endpoints PFS and OS. Predictive factors for AA and additional factors predictive of OS and PFS were not shared between cohorts. We hope that our findings will help in the development of prognosis scoring systems to help reduce-treatment related toxicities and accurately predict individual patient survival and progression times.

4.2 Introduction

For both patients and physicians, accurately predicting the prognosis once a malignancy has been diagnosed remains central to defining a treatment regimen that will increase the likelihood of achieving a positive tumour response while minimizing the risk of treatment related toxicities. Patients and their families are typically focused on the efficacy prognosis as they try to understand the seriousness of their illness and chances of medium to long-term survival. By contrast, many clinicians are interested in prognosis of treatment-related toxicity as it worsens patient quality of life and is often the cause of dose reduction or treatment discontinuation. In patients with advanced cancer, adverse events leading to hospitalisation are common and consequently represent a costly phenomenon associated with cancer care [158, 166-169]. Even in the absence of hospitalisation, if adverse events appear, healthcare providers typically offer a dose reduction or temporary treatment interruption in the hope that the adverse event may resolve and thereby allow continuation of treatment. In the event of severe or persistent adverse events, treatment discontinuation may be necessary. Any treatment interruption, whether temporary or permanent, has the potential to jeopardise the benefit from therapy. Chemotherapy induced adverse effects are therefore undesirable outcomes.

When making chemotherapy treatment plans, oncologists use many indicators to identify patients at risk for adverse effects. Performance status, a clinical estimate of functional status, is the most important of these indicators, and patients with poor performance status (e.g. an Eastern Cooperative Oncology Group (ECOG) [104] performance status >2) are generally considered to face more risks than benefits from chemotherapy. Beyond performance status, additional factors (including age and renal function) are also known to influence the risk of chemotherapy toxic effects [170, 171]. A number of previous studies have sought to develop discriminative approaches for assessing the risk of chemotherapy induced toxic responses [165, 171-173]. While these studies explored alternate patient populations, predictors, and toxic effect outcomes, they were all able to demonstrate that model-based approaches can identify toxicity risk factors and have the potential to improve risk stratification for chemotherapy toxic effects [168].

We set out to investigate and improve our understanding of the impact of baseline clinical and demographic patient characteristics on both efficacy and safety endpoints separately for patients

with lung cancer and patients with ovarian cancer. Specifically, we investigated the impact of several pre-treatment factors including age, sex, performance status (PS), renal function (RF), smoking status and ethnicity.

The ability to better identify patients at elevated risk for chemotherapy toxic effects has potential to improve patient outcomes at many levels. Specifically, better risk assessment for chemotherapy toxic effects would improve the chemotherapy informed-consent process, allow for modification of treatment regimens to reduce the risk of toxic effects, and identify patients who may benefit from aggressive therapy regimens around the time of chemotherapy initiation.

4.3 Methods

4.3.1 Patients

The current study was a retrospective review of data collected from patients with lung and ovarian cancers treated with platinum chemotherapy at Guy's and St. Thomas' NHS Foundation Trust. Baseline patient and demographic variables were extracted from standardised oncology outcome records completed at each clinic visit. A more complete description of these data is provided in Chapter 2.

4.4 Statistical analysis

Three endpoints (one safety and two efficacy) were examined for prediction by baseline variables: any-AE (AA), overall survival (OS) and progression free survival (PFS). Six baseline variables were included in the analyses: age, sex, performance status, renal function, ethnicity and smoking status. For the safety endpoint AA, the univariate prognostic potential of baseline variables was assessed using a binary logistic regression model in which each patient was classified as being either 'affected' if they experienced any severe adverse event (grade 3 or greater) at any cycle during the treatment phase, or 'unaffected' if they did not experience a severe adverse event during treatment.

For the survival endpoints (OS and PFS), the univariate prognostic potential of baseline variables was assessed using Cox regression models. Multivariable stepwise regression analysis with both

entry and removal levels as 0.05 was used to identify independent predictors for PFS, OS or AA from the covariates studied in the univariate analyses.

OS was defined as the time from first dose to death from any cause. PFS was defined as the time interval from first-dose to first disease progression or death from any cause if disease progression did not occur.

The baseline variables included in the models were: age, sex, performance status, renal function, ethnicity and smoking status. Prior to the analyses, we created clinically meaningful grouping of both renal function and performance status owing to the small number of patients which were category 3 for each variable. The grouping for both these variables therefore became “0-1” and “2-3”.

All analyses were performed separately in patients with lung cancer and patients with ovarian cancer. Statistical analyses were completed using SAS v.9.4 software.

4.5 Results

4.5.1 Patient characteristics

Data from 566 patients (223 ovarian patients and 343 lung patients) were included in this study. Table 2.4.1 summarises the demographic and clinical characteristics. In brief, the median age was 64 years. Excluding missing values, the overwhelming majority of patients were Caucasian (87.3%). Most patients had performance status in the 0-1 range (86.7%) and renal function scores in the 0-1 range (87%). The ovarian cohort had a higher proportion of patients with performance status 2-3 (19.4%) as compared with the lung cohort (9.7%). The proportion of patients in each renal function category was similar between cohorts.

4.5.2 Adverse events

The frequency of severe adverse events (CTCAE≥3) across all 4 cycles of therapy is presented by cohort and overall in Table 2.4.3

Across both cohorts, the number of severe events is greater than the number of patients indicating that some patients experienced multiple severe events. Figure 4.5.1 presents the patient frequency against number of severe adverse events by cohort. 264 patients (47%) had 0 (no)

severe adverse events; 157 (28%) had one severe adverse events; 83 (15%) had two severe adverse events; 34 (6%) had 3 severe adverse events, 14 (2.5%) had 4 severe adverse events, 9 (1.6%) had 5 severe adverse events, 4 (0.7%) patients had 6 severe adverse events, and 1 (0.17%) patient had 7 severe adverse events.

The proportion of patients experiencing a singular or multiple severe adverse events is similar between cohorts (Figure 4.5.1).

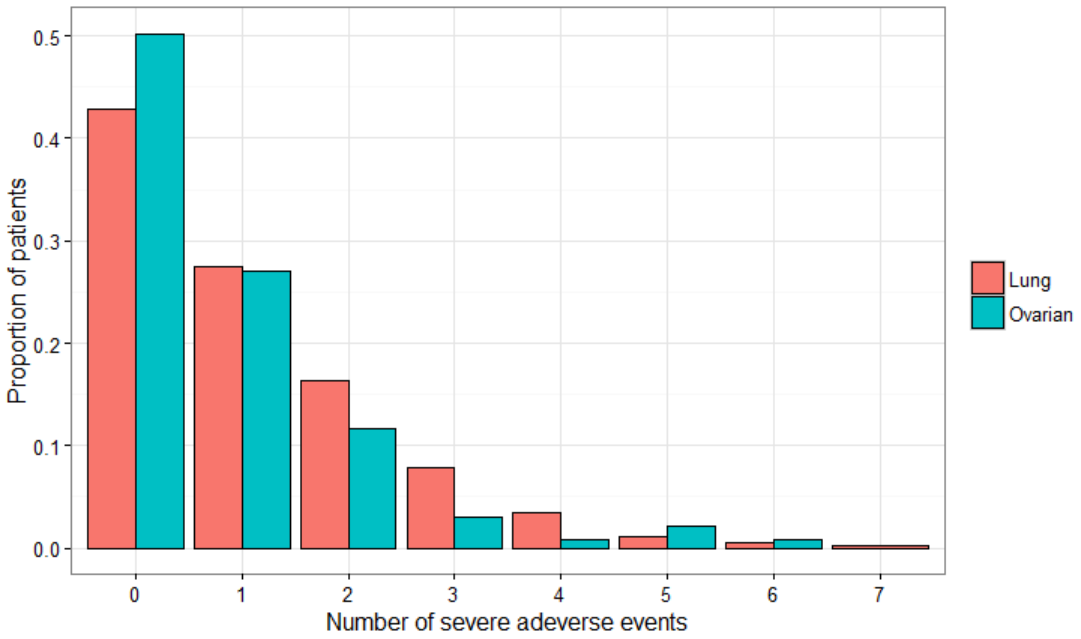


Figure 4.5.1 Proportion of patients against number of severe adverse events by cohort and overall

4.5.3 Prognostic value of patient characteristics – safety endpoint

Among baseline patient characteristics, univariate analysis showed that in the ovarian cohort, performance status (p=0.0276) and renal function (p=0.0272) were associated with patient risk of experiencing a severe adverse event. In the lung cohort, age (p=0.0045), sex (p=0.0076) and renal function (p=0.0281) were all associated with patient risk of experiencing a severe adverse event (Table 4.5.1.)

When fitted together in a multivariable analysis the only baseline characteristic that remained significantly associated with risk of severe adverse event in the ovarian cohort was performance

status (p=0.0438: p-value from Type III analysis of effects[1]); where patients with PS 2-3 were more likely to experience a toxic event compared with patients who were PS 0-1 (OR=2.297 95%CI=1.096-4.813, p=0.0125, AUC=0.563). In the lung cohort only age (p=0.0042) and sex (p=0.0070) remained significant in the multivariable analysis with a combined AUC of 0.62 (Figure 4.5.2). Specifically, patients who were older were more likely to experience a severe adverse event (OR=1.032, 95%CI=1.010-1.055 p=0.0042) and men were less likely to experience a severe adverse event as compared with women (OR=0.540 95%CI=0.345-0.845, p=0.0070)

4.5.4 Prognostic value of patient characteristics – efficacy endpoints[2]

In the ovarian cohort, of the examined baseline characteristics, performance status, renal function, age and smoking status were predictive of both OS (p<.0001, p=0.0006, p=0.0008, and p=0.0418 for performance status, renal function, age and smoking status respectively) (see Table 4.5.2) and PFS (p<.0001, p=0.0063, p=0.0335, p=0.025 and p=0.1544 for performance status, renal function, age and smoking status respectively) (see Table 4.5.3). When all factors were entered into a multivariable analysis, both performance status (p<.0001) and renal function (p=0.0145) were identified as independent predictive factors for overall survival. Following multivariable analysis, performance status was the only independent factor of progression free survival (p<.0001). Across both variables (OS and PFS), patients who were category 0-1 had improved median survival compared with patients of category 2-3 (Figure 4.5.3).

In the lung cohort, univariate analysis identified sex (p=0.0308) and performance status (p= 0.0016) as predictive of OS (see Table 4.5.2) and performance status (p= 0.0012) and smoking status (p=0.0191) as predictors of PFS (see Table 4.5.3.). Multivariable analysis, confirmed both sex (p=0.0448) and performance status as independent prognostic factors of overall survival (Table 4.5.4). Similarly, results from the multivariable analysis of PFS identified smoking status

---

[1] Type III analysis of effects - test for the significance of each explanatory variable, under the assumption that all other variables entered in the model equation are present. A significant p-value indicates that at least one of the subcategories of the predictor variable is associated with the outcome. In the univariate analyses we excluded 'missing' status from the predictor variable but left it in for the multivariable analyses as exclusion would have resulted in decreased sample size across predictors.

[2] The text presents the Type III test of overall effect for each prognostic variable. Tables 4.5.4-4.5.6 present the OR/HR and corresponding p-value for each predictor level relative to the reference level.

(p=0.0171) and performance status (p= 0.0009) as independent predictive factors for progression free survival (see Table 4.5.4).

Table 4.5.1 Univariate analysis of potential prognostic factors for any-AE (AA)

Characteristic	***Ovarian***				***Lung***			
	Affected†	***Unaffected***	***OR (95% CI)***	***p-value***	***Affected***	***Unaffected***	***OR (95% CI)***	***p-value***
Age	.	.	1.016 (0.995-1.037)	0.1332	.	.	1.032 (1.010-1.054)	0.0045
Sex	116	107		.	145	198		.
Female	116	107		.	50	97	reference	.
Male	.	.	NE‡	.	95	101	0.548 (0.352-0.852)	0.0076
ECOG Performance status	116	107		.	145	198		.
0-1	86	72	reference	.	131	168	reference	.
2-3	13	25	2.297 (1.096-4.813)	0.0276	9	23	1.992 (0.892-4.449)	0.0929
Renal Function	116	107		.	145	198		.
0-1	103	84	reference	.	127	160	reference	.
2-3	9	19	2.588 (1.113-6.018)	0.0272	13	35	2.137 (1.085-4.209)	0.0281
Ethnicity	116	107		.	145	198		.
Caucasian	101	88	reference	.	122	176	reference	.
Asian	1	2	2.295 (0.205-25.748)	0.9756	2	4	1.386 (0.250-7.688)	0.5090
Black	10	8	0.918 (0.347-2.428)	0.9645	13	12	0.640 (0.282-1.450)	0.5500
Other	.	5	NE ᵀ	-	6	2	0.231 (0.046-1.164)	0.0718
Mixed	1	1	1.148 (0.071-18.621)	0.9673	1	3	2.080 (0.214-20.228)	0.3458
Smoking	116	107		.	145	198		.
Never	33	31	reference	.	17	24	reference	.
Ex-smoker	6	8	1.419 (0.442-4.557)	0.2547	44	53	0.853 (0.408-1.786)	0.5290
Current	22	11	0.532 (0.222-1.276)	0.0894	77	111	1.021 (0.514-2.028)	0.6759

† Affected patients are those that experienced a severe (Grade 3 or greater) adverse event

‡NE – not able to estimate as the ovarian cohort contained no males

ᵀ NE – Not able to estimate due to there being no controls within this category of ethnicity

Table 4.5.2 Univariate analysis of potential prognostic factors for overall survival

Characteristic†	***Ovarian***				***Lung***			
	Events	***Censors***	***HR (95% CI)***	***p-value***	***Events***	***Censors***	***HR (95% CI)***	***p-value***
Age	70	153	1.039 (1.016-1.063)	0.0008	221	122	1.011 (0.997-1.025)	0.1274
Sex	70	153		.	221	122		.
Female	70	153		.	84	63	reference	.
Male	.	.	NE‡	.	137	59	1.350 (1.028-1.772)	0.0308
ECOG Performance status	70	153		.	221	122		.
0-1	42	116	reference	.	188	111	reference	.
2-3	22	16	4.050 (2.388-6.869)	<.0001	25	7	1.823 (1.199-2.772)	0.0050
Renal Function	70	153		.	221	122		.
0-1	54	133	reference	.	181	106	reference	.
2-3	14	14	2.855 (1.560-5.226)	0.0007	35	13	1.236 (0.860-1.776)	0.2515
Ethnicity	70	153		.	221	122		.
Caucasian	55	134	reference	.	196	102	reference	.
Asian	1	2	1.443 (0.199-10.469)	0.7171	4	2	1.114 (0.414-3.002)	0.8305
Black	8	10	1.525 (0.723-3.218)	0.2681	17	8	1.051 (0.640-1.726)	0.8437
Other	3	2	2.416 (0.751-7.777)	0.1391	2	6	0.343 (0.085-1.382)	0.1324
Mixed	1	1	1.019 (0.139-7.460)	0.9854	1	3	0.416 (0.058-2.972)	0.3823
Smoking	70	153		.	221	122		.
Never	25	39	reference	.	27	14	reference	.
Ex-smoker	4	10	0.690 (0.240-1.986)	0.4917	65	32	0.982 (0.627-1.538)	0.9360
Current	10	23	0.863 (0.413-1.801)	0.6940	118	70	0.946 (0.622-1.437)	0.7932

† 'Missing' status within each characteristic excluded from the analysis

‡NE – not able to estimate as the ovarian cohort contained no males

Table 4.5.3 Univariate analysis of potential prognostic factors for progression free survival

Characteristic†	***Ovarian***				***Lung***			
	Events	***Censors***	***HR (95% CI)***	***p-value***	***Events***	***Censors***	***HR (95% CI)***	***p-value***
Age	148	65	1.017 (1.001-1.032)	0.0335	275	62	1.000 (0.988-1.012)	0.9664
Sex	148	65		.	275	62		.
Female	148	65		.	114	30	reference	.
Male	.	.	NE‡	.	161	32	1.092 (0.858-1.388)	0.4754
ECOG Performance status	148	65		.	275	62		.
0-1	93	60	reference	.	234	59	reference	.
2-3	32	3	3.305 (2.184-5.001)	<.0001	30	2	1.946 (1.325-2.857)	0.0007
Renal Function	148	65		.	275	62		.
0-1	118	60	reference	.	229	52	reference	.
2-3	24	4	1.996 (1.282-3.110)	0.0022	39	9	0.975 (0.693-1.371)	0.8823
Ethnicity	148	65		.	275	62		.
Caucasian	128	55	reference	.	241	52	reference	.
Asian	2	1	0.982 (0.242-3.981)	0.9801	5	1	1.564 (0.644-3.798)	0.3236
Black	12	4	0.994 (0.548-1.802)	0.9845	20	5	0.989 (0.627-1.561)	0.9620
Other	3	2	0.882 (0.278-2.795)	0.8312	6	2	0.784 (0.349-1.763)	0.5564
Mixed	1	1	0.306 (0.042-2.212)	0.2404	2	1	0.613 (0.152-2.467)	0.4905
Smoking	148	65		.	275	62		.
Never	46	15	reference	.	37	4	reference	.
Ex-smoker	10	4	0.807 (0.407-1.600)	0.5382	76	19	0.555 (0.372-0.827)	0.0038
Current	21	11	0.873 (0.521-1.464)	0.6067	150	36	0.614 (0.427-0.884)	0.0087

† 'Missing' status excluded from the analysis

‡NE – not able to estimate as the ovarian cohort contained no males

Table 4.5.4 Results of multivariable analysis - independent prognostic factors for OS and PFS identified by Cox multivariable proportional hazard model with stepwise selection; independent prognostic factors for AA identified by multivariable logistic regression model with stepwise selection.

		Ovarian			***Lung***		
Parameter	***Reference***	***Hazard/Odds Ratio***	***95% Hazard Ratio Confidence Limits***	***p-value***	***Hazard/Odds Ratio***	***95% Hazard Ratio Confidence Limits***	***p-value***
Any AE (AA)							
Age			NA†		1.032	1.010- 1.055	0.0042
Sex	Male vs. Female				0.540	0.345- 0.845	0.0070
Performance status	2-3 vs. 0-1	2.297	1.096 - 4.813	0.0125		NA†	
Overall Survival (OS)							
Sex	Male vs. Female				1.322	1.006-1.736	0.0448
Performance status	2-3 vs. 0-1	3.830	2.243- 6.540	<.0001	1.770	1.163-2.694	0.0077
Renal Function	2-3 vs. 0-1	2.144	1.164- 3.950	0.0145		NA†	
Progression Free Survival (PFS)							
Smoking Status	Current vs. Never		NA†		0.598	0.415 - 0.861	0.0057
	Ex-Smoker vs. Never				0.568	0.381 - 0.847	0.0056
Performance status	2-3 vs. 0-1	3.305	2.184 - 5.001	<.0001	2.063	1.389 - 3.064	0.0003

† - NA – Not Applicable as the variable was not independently prognostic within the cohort

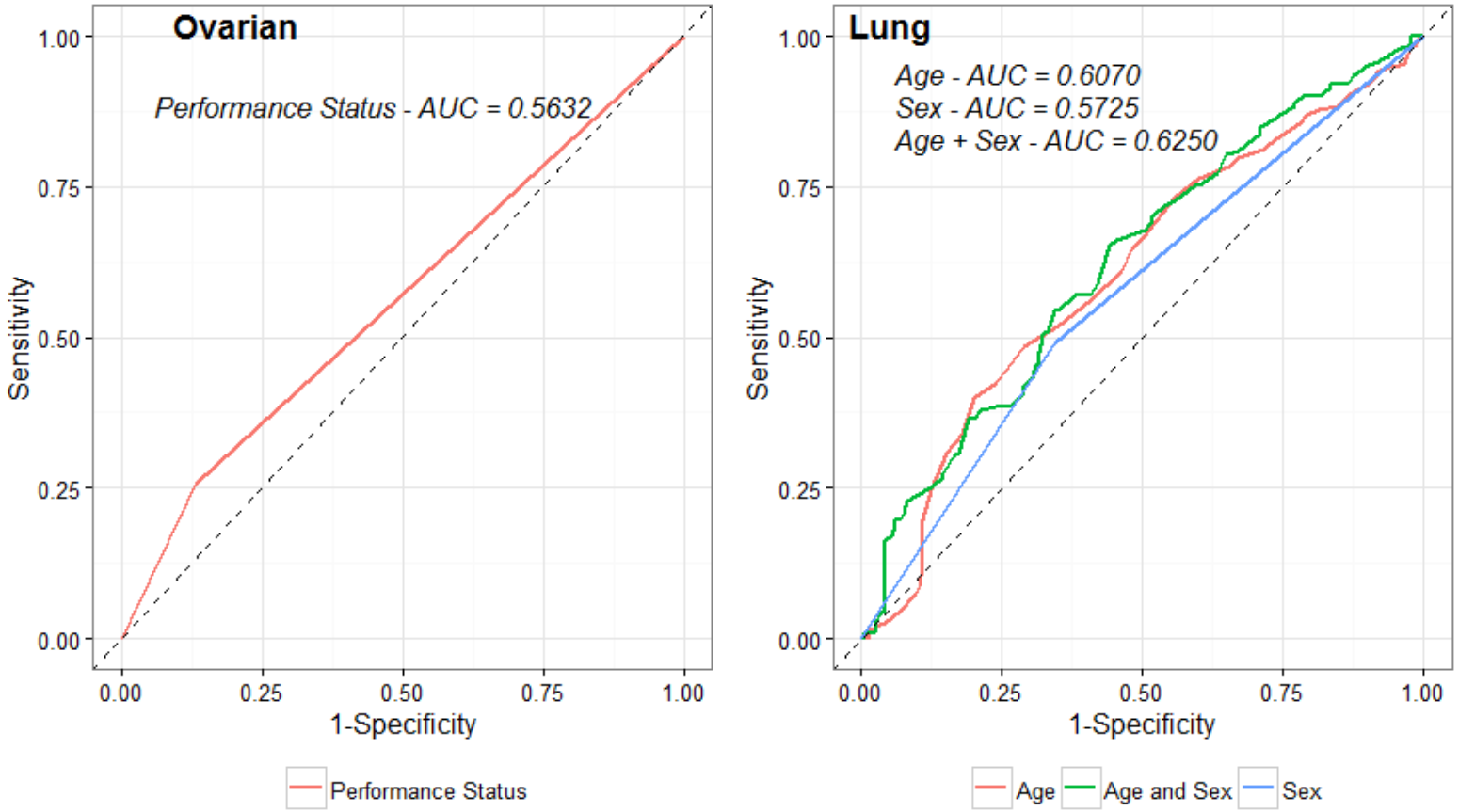


Figure 4.5.2 Results of multivariable analysis by cohort.
In the ovarian cohort, only performance status was a significant predictor of AA. In the lung cohort, age and sex were both predictors of AA. The lung plot shows the AUC associated with each prognostic variable from the univariate analysis and the combined AUC from the multivariable analysis.

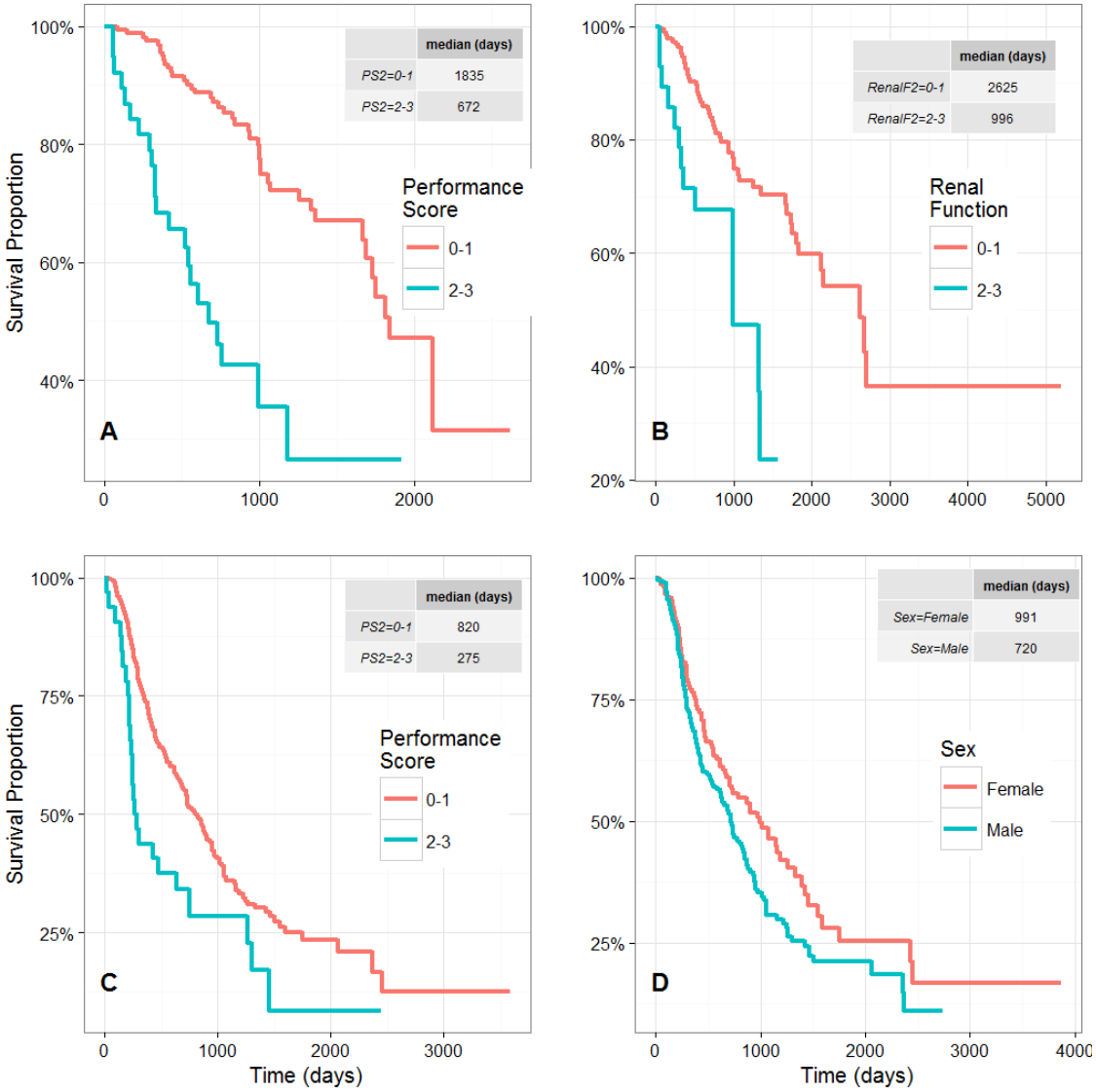


Figure 4.5.3 Multivariable analysis - independent prognostic factors for OS.
(A) Ovarian Cohort – performance status (PS); (B) Ovarian Cohort – renal function (RF); (C) Lung cohort – performance status (PS); and (D) Lung cohort – sex. The grey shaded table in each plot panel shows the median OS by strata.

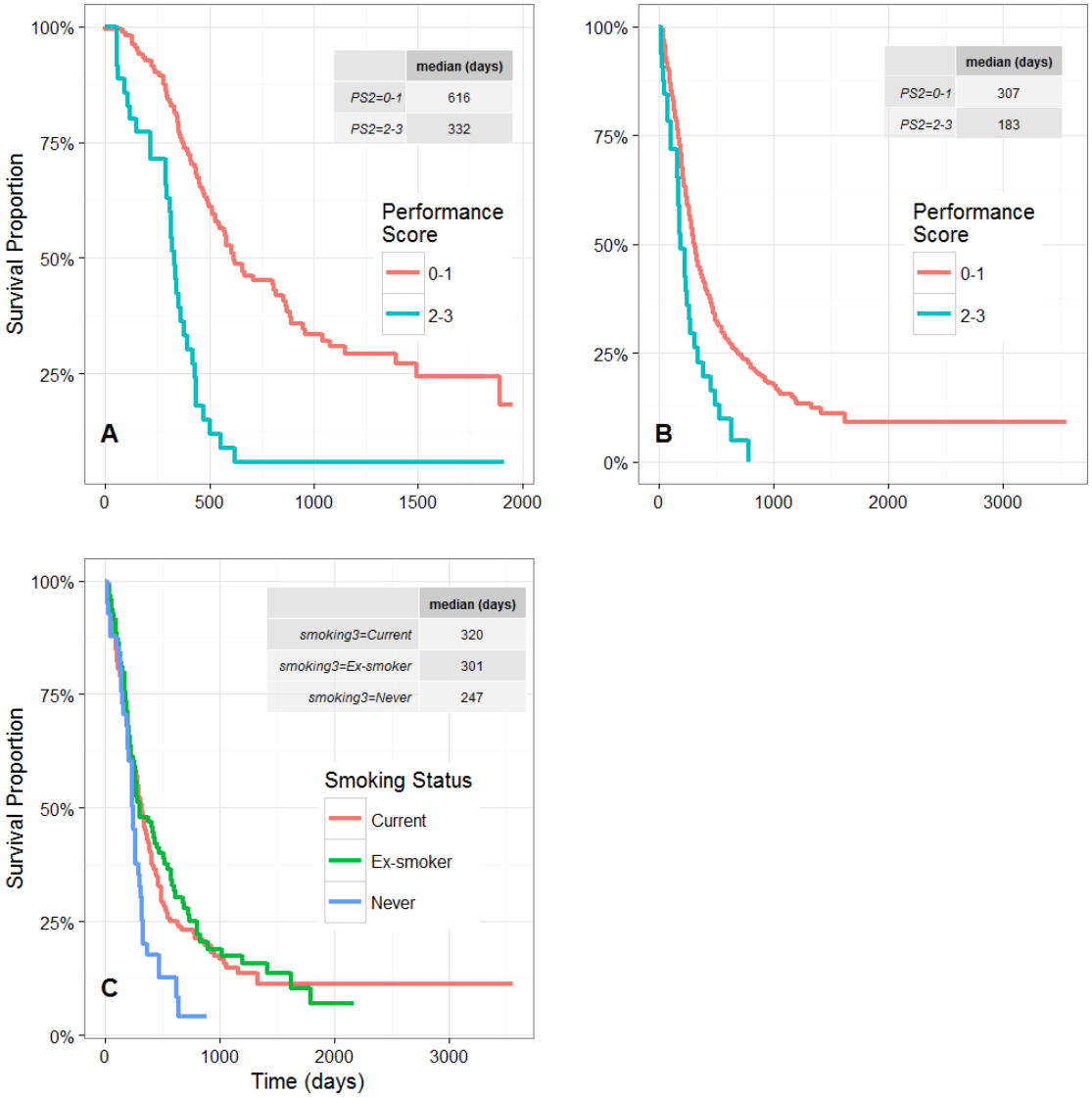


Figure 4.5.4 Multivariable analysis - independent prognostic factors for PFS.
(A) Ovarian Cohort – performance status (PS); (B) Ovarian Cohort – renal function (RF); (C) Lung cohort – performance status (PS); and (D) Lung cohort – sex. The grey shaded table in each plot panel shows the median PFS by strata

4.6 Discussion

In this study, we have tested patient characteristics for their prognostic value in relation to both safety and efficacy outcomes in response to platinum chemotherapy. This study identified that a low performance status predicted improved overall survival and progression free survival in both lung and ovarian cancer patients. In addition, low performance status was associated with reduced risk of experiencing a severe adverse event in the ovarian cohort. These results support previous studies which identified performance status as a significant prognostic factor in small cell lung cancer [174, 175], non-small cell lung cancer [176] and ovarian cancer [177]. In the ovarian cohort, renal function was also identified as an independent prognostic factor for OS. This result was unexpected as clinicians typically use renal function when generating a performance status [178], resulting in a strong correlation between the two measures. Impaired renal function is known to be associated with decreased physical function and physical activity [179-181] and several previous studies have identified an association between renal function and physical performance [182-184]. Our results suggest that renal function scores offer additional information that helps predict overall survival beyond the impact upon physical activity.

In the lung cohort, age and sex were identified as prognostic of safety but were not associated with either survival endpoint. This contrasts with previous studies which have identified old age at diagnosis as an important determinant of survival prognosis [7, 185-187]. In general, patients are more likely to exhibit co-morbidities with increasing age in patients with lung cancer [188, 189]. Consequently, the use of aggressive chemotherapy in elderly patients is controversial because of the increased possibility of adverse event side effects compared with younger patients [190]. Our finding that increased age was associated with an increase in the incidence of adverse events in the lung cohort supports the perspective that older age increases the risk of potentially harmful adverse events and therefore patient age needs to be accounted for when deciding upon a treatment regime.

Smoking is a recognised risk factor for lung cancer [191-193], and we had hypothesised that current and ex-smokers would have reduced PFS and OS as compared with never-smokers. It was surprising that current, ex-smokers and unknown smoking status patients all had a longer progression free survival as compared with never-smokers. Ignoring patient error in reporting

accurate smoking status, it could be that never-smokers who develop lung cancer have been exposed to cancer risk agents that result in a more aggressive disease state. As the current study did not collect disease staging, it is impossible to test this hypothesis but we believe this feature warrants further research. We did not detect an association between smoking status and survival. This may be due to a lack of power in the OS analysis (PFS has more events) but it might also imply that there are post-progression survival differences between patients of alternate smoking status classes.

We observed that male patients in the lung cohort were less likely to experience an adverse event as compared with females, however, male patients had reduced survival as compared with female patients. This is consistent with previous reports that have documented an OS [194, 195] and PFS [196] advantage in women treated for lung cancer; and also greater toxicity for females as compared with males in response to chemotherapy [197-200]. Whether the survival difference results from improved treatment response or different tumour biology (including stage and subtype) is not well understood [195].

The strength of the present study is that it represents essentially unselected patients at a cancer treatment centre within a major metropolitan area. Unlike clinical trials, there were no demographic or baseline variable exclusions (i.e. age, PS, renal function) in the sample population. These data are therefore representative of the patient heterogeneity that clinicians are presented with in regular oncology patient care.

Several of the baseline characteristics we tested did not show a significant prognostic value in the multivariable models assessing PFS and OS. These markers have been explored in the previous clinical studies with varying results. The discrepancy between their results and our may be explained by differences in the disease stage of the patient population, lack of power due to small numbers of patients, or the use of different treatment schedules. Additional variables of dose-intensity, concomitant medications and disease staging would be valuable to test for any confounding. For consistency with the other sections of this work, we tested for an association between baseline characteristics and neutropenia and gastrointestinal disorders (see 0: Table 12.2.1 and Table 12.2.2). These analyses of adverse event subgroups failed to replicate all of the baseline characteristic-adverse event associations observed in the Any-AE analysis. This is

most likely a power issue as the number of Grade 3 or greater events is much smaller when we slice the severe adverse event pool by different subtypes.

Lastly, we observed differences in the prognostic characteristics between cohorts but cannot explain the discrepancy. Both cohorts had a similar proportion of patients experiencing a severe event and a similar distribution of patients with no adverse event, a single event and multiple events. The key difference between cohorts is the survival distribution. If we assume that both cohorts have the same underlying prognostic factors, then it could be argued that the aggressive nature of lung cancer (shorter survival) reduces our ability to detect factors that will ameliorate or accelerate hazard, and therefore the ovarian data findings are more relevant. Alternatively, the longer survival time associated with ovarian cancer as compared with lung cancer means that the survival curves are subject to bias because of confounding associated with competing risks, concomitant therapy and subsequent therapy. This suggests that the results from the associations with the lung cohort could be viewed as unbiased and therefore more clinically relevant. It is also possible that ovarian and lung cancer do not share the same underlying prognostic factors. The cancer characteristics could be disparate enough to render the factors prognostic for safety and efficacy different between each cancer type.

Our study emphasises the importance of performance status assessment in patients with ovarian and lung cancers as the score provides useful prognostic information. In advanced ovarian cancer, renal function should be assessed in addition to performance status, as this information is of importance in influencing treatment outcomes. In lung cancer patients receiving platinum chemotherapy, we confirm the importance of gender in both the risk of developing a toxic response and in its prognosis of survival.

The divergent results of our current and previous studies highlight the complexity of patient prognosis and represent a barrier to the use of these factors in daily practice. Establishing clear prognostic variables for efficacy and safety may help clinicians to decide on a patient specific treatment regime that effectively balances benefit and risk.

Our work considered each outcome (OS, PFS and AA) separately, a common way of analysing several outcomes from the same trial [201]. An alternative analysis strategy could have been to analyse each predictor against the set of outcomes through one of several statistical methods

designated as multivariate methods [202]. In the case where we considered a multivariate model for the survival endpoints only (OS and PFS) there are 3 alternate multivariate strategies: the joint survival function [203], the marginal survival function [204] and the conditional survival function [205]. Alternatively, if we were to ignore the survival times and categorise each according to their status at the end of the observation period as either having experienced the survival event or not (binary) then a multivariate logistic model could be used to explore both survival outcomes and the safety outcome simultaneously. Multivariate analysis is frequently used to control type-I error probability, the logic being that if the multivariate analysis yields significance then it may be appropriate to carry out multiple univariate tests to identify the specific outcome variable that associates with a given predictor or set of predictors [206].

Huberty and Morris [206] outline four instances (within the context of ANOVA) in which multiple univariate analyses may be favourable to a multivariate analysis: 1) when the outcome variables are conceptually independent, that is the question of interest is how the predictor variable(s) affect each of the outcome variables without interest in the relationship between outcome variables; 2) when the research being conducted is exploratory in nature, i.e. a new predictor and/or outcome are being studied and the effects of the predictor upon the outcome are being investigated to reach “tentative conclusions”; 3) when the outcome variables in the current study have previously been studied within univariate context; 4) and lastly when the research question is needed to show that two or outcomes are equivalent with respect to a set of predictors. For our study, the endpoints were not conceptually independent; OS is frequently considered conditional upon PFS, as patients do not tend to die of non-progressing tumours. In support of the multiple analyses, our work was both exploratory and followed numerous previous studies which have explored survival and safety outcomes independently. Additionally, analysing the outcomes separately, as we did, does not require that the outcomes are measured on the same scale, and the difference in the recording of date between survival and safety events within this data lends itself towards separating survival and safety outcomes. However, through analysing both safety and binary survival outcomes together, correlation between the outcomes can be used to increase power to detect predictor effects. However, the gain in power from the multivariate model must be considered against the added number of predictor terms. As discussed further in Chapter 5, within the logistic regression framework, increasing the number of covariates inflates both the parameter estimate and the variance of the estimate for the variable of interest [207, 208], so

while multivariate models are generally considered more powerful, this may be offset through the additional predictor term(s) required in the model.

Chapter 5. Genetic variation as predictor of efficacy and toxicity in platinating agent treated patients

If it were not for the great variability among individuals, medicine might as well be a science and not an art.

– Sir William Osler

5.1 Abstract

Background

Lung and ovarian carcinomas are common malignancies and each is a leading contributor to cancer deaths worldwide each year. Both forms of cancer are treated with chemotherapy regimens involving a 'platinum' compound.
Previous studies have identified that clinical factors (including age, sex, performance status and disease stage), and the genetic profile of an individual are each independent predictors of efficacy response and toxicity in platinum treated patients.
Most genetic association studies only used the genomic data without adjusting for established clinical covariates that are known to have predictive value.
The aims of the present study were thus:

1) To explore how combining prognostic baseline factors (age and sex) with GWAS analysis would alter the pattern and strength of observed SNP-phenotype association results.
2) To identify novel candidate single nucleotide polymorphisms (SNP) associated with overall survival (OS), progression free survival (PFS), neutropenia (NEU), or gastrointestinal disorder (GID) in lung and ovarian cancer patients.

Methods

An exome array of approximately 30,000 SNPs was tested for association against each of the safety and efficacy phenotypes in both a lung (N=343) and ovarian cancer cohort (N=223).
Association of the SNPs with OS and PFS was determined by using the Cox regression model. Association of the SNPs with NEU and GID was determined by using the logistic regression model. We analysed the association of each SNP with and without the inclusion of the clinical variables age and sex.

Results

Of the ~30000 successfully genotyped SNPs, 2 SNPs passed multiple correction for association with overall survival in the lung cohort (rs17117678 with in the *OMA1* gene and rs35075952 within the *TACSTD2* gene) with and without the addition of clinical covariates. No other multiple testing corrected associations were significant. Both SNPs had reduced association signals once modelled with baseline covariates. Across all the examined outcomes (OS, PFS, NEU and GID), the addition of baseline covariates either reduced the strength of association or had a negligible effect.

Conclusions

Our results indicate that further work needs to be done to explore how best to model the effect of clinical and genetic sources of variation. Our study identified SNPs in key candidate genes associated with overall survival in lung cancer. If validated in a replication cohort, these findings might provide opportunities to personalise therapeutic strategies.

5.2 Introduction

Platinum-based doublet therapy including cisplatin or carboplatin combined with taxanes, gemcitabine, vinorelbine, and etoposide are standard first-line treatments for advanced non-small cell lung cancer (NSCLC), small cell lung cancer [209-211] and ovarian cancer [212, 213].

The use of platinum therapy is not without risk, the potential antineoplastic benefit of these frequently prescribed drugs is often compromised by severe side effects including gastrointestinal upset and neutropenia. Cisplatin is frequently regarded as the platinum drug with the most severe side effects, including nausea and vomiting, neurotoxicity, nephrotoxicity (kidney damage), ototoxicity (hearing damage) [25, 54-56]. In response to the unwanted side effects, known as adverse events (AEs), therapy may be either continued at a reduced dose or discontinued completely [214, 215]. Both courses of action potentially compromise patient benefit as patients are denied the full therapeutic benefit of the platinum agent.

There is considerable inter-individual variation in the efficacy and safety responses to platinum therapy [215]. The inter-individual variation can likely be explained, in part, by genetic variation [216, 217] as well as baseline and clinical prognostic factors [218]. Response prediction from genetic data is an active field in today's medical research. Most often association analyses utilise only the genomic data without including established clinical covariates that often are known to have predictive value with respect to the phenotype. Correspondingly, it might be beneficial to also integrate complementary information from patients when building multivariable risk prediction models for a clinical endpoint, such as treatment response or survival [219]. Age adjustment is widely used in studies comparing survival of different cancer patient populations [220]. This is because patient age is both important risk factor but also prognostic of survival for many kinds of cancer [221] (including lung cancer [222] and ovarian cancer [223, 224]). Patient sex has also been identified as a predictor for lung cancer survival [225]; with women having better survival than men across all cancer stages and independent form the therapeutic approach [226].

Several recent papers have included both clinical and genetic covariates when seeking genetic associations with cardiomyopathy [227], multiple sclerosis [228],and cancer [229]. The integration of clinical and genetic has the potential to improve our understanding, and our ability to predict disease risk and prognosis [229, 230]. Beyond accuracy of the predicted outcome,

conditioning on covariates associated with phenotype can increase study power by reducing phenotypic variance [231]. However it is well known that the addition of covariates can result in a dramatic loss in power depending on the type of model used to ascertain the SNP-outcome association [207, 208]. With respect to survival outcomes analysed with a Cox regression model, it is not obvious how to the addition of clinical covariates may impact our ability to ascertain SNP-outcome associations [232]. The inclusion of covariates within a Cox model can both improve the accuracy of the predictor of interest and/or help with achieving proportional hazards in which case their inclusion is necessary to meet the model assumptions. Our analyses focus on the former however situation we accept that the inclusion of covariates in alternate studies may be for a different reason.

Consequently, we sought to examine if combining clinical and genomic information influenced the strength of SNP-phenotype association signals.

The objectives of this paper are thus:

1) To explore how combining prognostic clinical factors with GWAS analysis would alter the pattern and strength of observed association results.
2) To identify novel candidate single nucleotide polymorphisms (SNPs) associated with overall survival (OS), progression free survival (PFS), neutropenia (NEU), or gastrointestinal disorder (GID) in lung and ovarian cancer patients

5.3 Methods

5.3.1 Study population

A more detailed description of the population has been published elsewhere [60] and is provided in the preceding chapters. In brief, all patients were recruited from within Guy's and St. Thomas' Hospitals NHS Foundation Trust. Patients who began treatment for either lung or ovarian cancer between 02/17/2000 and 05/17/2013 were identified by the institutional treatment registry. Of these, 589 patients were treated with first-line platinating-agent based chemotherapy. Patients eligible for possible inclusion into the study were approached in order to obtain consent for a laboratory blood draw to be used in genotyping. Information regarding patient demographics, baseline characteristics, and outcomes was collected through manual review of the patient electronic records. The institutional ethics committee approved the study.

5.3.2 Genotyping

DNA was extracted from ethylenediaminetetraacetic acid (EDTA) whole blood using the QIAamp DNA Blood Mini Kit (Qiagen, Crawley, West Sussex, UK). All patients were genotyped with the Illumina Infinium HumanExome BeadChip v1.1 (Illumina, San Diego, CA, USA), which provides coverage of over 240,000 SNPs, including functional exonic variants (>90%), disease-associated tag markers from recently published GWAS, ancestry-informative markers, and other markers. Genotypes were called using GenCall Data Analysis Software v1.0 clustering algorithm in Genome Studio v2011.1 (Illumina). Genotyping was conducted at the King's Genomic Centre. Genotyping and calling was performed on this study prior to my involvement, and I obtained genotype and phenotype data from Dr. Adele Corrigan on 13th October 2015.

Several quality-control filters were applied to samples and variants prior to analysis. Individuals were excluded for the following reasons: sex mismatch, individual call rate of <97%, SNP call rate of <99%, Hardy–Weinberg Equilibrium *p*-value $<10^{-6}$, cryptic relatedness and minor allele frequency <0.05. Population stratification was assessed using principal components. Principal components analysis was carried out using EIGENSOFT on all common SNPs (minor allele frequency 0.01). All quality control and principal components analysis was completed by Dr Jemma Walker (Statistical Genetics - King's College London).

After exclusions, a total of 29328 variants and 343 individuals were available for analysis from the lung cancer cohort and a total of 29028 variants and 223 individuals were available for analysis from the ovarian cancer cohort.

5.3.3 Statistical methods

5.3.3.1 Sample size

The data for this study were originally collected for an examination of compound safety events. Consequently, the study was not designed for the analyses we present in this paper. We computed post hoc power analysis using two approaches depending upon the endpoint: 1) for survival endpoints we use a modification of the sample size estimation methods for non-binary covariates [233] which is itself an extension of an earlier formula for the case of a single, binary covariate derived by Schoenfeld [234]; 2) for safety endpoints we computed the effective sample size and statistical power using the case-control for discrete traits option in the web browser

program, Genetic Power Calculator developed by Purcell et al., (http://pngu.mgh.harvard.edu/~purcell/gpc/).

Both types of power calculations were conducted under the assumption of an additive genetic model with a minor/risk allele frequency (MAF) of 0.1 and 0.2. These thresholds were chosen as representative of common variant frequencies within complex disease, to place in context the power of our datasets.

The Genetic Power Calculator (GPC) [235] calculates the power for a genetic association study. It requires input parameters of the minor allele frequency and the odds ratio (OR) for risk conferred by the allele, in addition to sample size and significance level. GPC uses the OR and population allele frequency to calculate the allele frequency in patients with and without adverse events, and then calculates the power to detect this difference in allele frequencies between the two groups.

5.3.3.2 Significance threshold

Our study analysed ~30,000 SNPs across 4 phenotypes (2 efficacy endpoints and 2 safety endpoints). To maintain an overall error rate of 5%, applying the principles of Bonferroni correction, we define our significance threshold for evidence of association as:

$$\frac{\alpha}{(no.of\ phenotypes) \times (no.of\ SNPs)} = \frac{0.05}{4 \times 30{,}000} \cong 2.1\mathrm{e}-7 \qquad (5.3.1)$$

5.3.3.3 Safety endpoints

Based upon the observed number of patients experiencing neutropenic or gastrointestinal adverse events in each cohort we calculated that the prevalence for neutropenia was 0.27 in the ovarian cohort and 0.35 in the lung cohort. For gastrointestinal events, the prevalence was 0.09 for the ovarian cohort and 0.13 for the lung cohort. Using the calculated prevalences and assuming a risk allele frequency of 0.2, Table 5.3.1 presents the odd ratio (OR) that we had 80% power to detect in each cohort for each safety endpoint:

Table 5.3.1 Minimum OR (conferred by each copy of the risk allele) needed to provide 80% power to detect an association between a SNP the safety phenotype based upon the observed prevalence of cases assuming a frequency of 0.1 or 0.2 for the risk allele (MAF) in the population ($\alpha = 2.1e-7$)

	MAF	***Ovarian***	***Lung***
Neutropenia	0.1	7.4	5.0
	0.2	4.8	3.4
Gastrointestinal Disorder	0.1	9.9	5.6
	0.2	8.7	4.3

5.3.3.4 Efficacy endpoints

For survival analyses, the power is a function of the number of events rather than the sample size. With two equally sized comparator groups, for a two-sided test of α, the number of events required to achieve power of 1-β for a specific hazard ratio (Δ) is given by [234]:

$$D = \frac{4\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{(\log \Delta)^2} \qquad (5.3.2)$$

Where $Z_{\alpha/2}$ and Z_{β} are standard normal percentiles.

If, however the fraction of patients is not equal between comparator groups then (5.3.2)) will underestimate power. A modified equation taking into account the relative proportion of each comparator arm is [236]:

$$D = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{(\log \Delta)^2} \times \frac{1}{\pi_1 \pi_2} \qquad (5.3.3)$$

Where $\pi_1$and π_2 are the proportions to be allocated to each of the two comparator groups.

When the group allocation is equal (i.e. $\pi_1 = \pi_2 = 0.5$) then (5.3.2) and (5.3.3) will yield the same result.

It is interesting to note that in both (5.3.2) and (5.3.3) the number of censored observations does not enter in to the power calculations. To obtain a formula for the sample size, *N*, we need to

inflate *D* by dividing by P_E, the proportion of subjects expected to fail during the observation period of the study.

Thus, the formula for N is [236]:

$$N = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{(\log \Delta)^2} \times \frac{1}{\pi_1 \pi_2 P_E} \tag{5.3.4}$$

All of the above equations assume that there are only two groups undergoing comparison (a binary covariate). For a large proportion of the examined SNPs, patients will fall into one of *three* categories (homozygous major allele, heterozygous, and homozygous minor/risk allele).

Hsieh and Lavori [233] provides a method for calculating the required number of events when the covariate X_1 has more than 2 levels (i.e. is non-binary). The authors demonstrate that in a univariate model, without making assumptions about the distributions of covariate X_1, the total number of deaths required is given by the following formula:

$$D = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\sigma^2 (\log \Delta)^2} \tag{5.3.5}$$

Where σ^2 is the variance of X_1. Equation (5.3.5) is similar to equation (5.3.3) except that the variance of X_1, $\pi_1 \pi_2$ has been replaced by a more general term, σ^2.

If X_1 is binary ($X_1 = 0$ or $X_1 = 1$) the variance (σ^2) of X_1 (using the sample variance calculation rather than a population variance calculation) will be equal to 0.5 and equation (5.3.5) will yield an identical result to equation (5.3.2).

While the authors do not explicitly state in their article, implicit in equation (5.3.5) is that the group sizes for each level of X_1 are equal.

If X_1 is made up of 3 levels (AA, Aa, aa) then the variance of X_1 will be 1. Using similar logic for the link between equations (5.3.2) and (5.3.3), we surmise that:

$$D = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{(\log \Delta)^2} \times \frac{(\frac{1}{3})^3}{\pi_1 \pi_2 \pi_3} \tag{5.3.6}$$

Which means that equations (5.3.6) and (5.3.5) will produce identical results when there are an equal number of patients in each of the genotype groups.

Again using similar logic to the link between equations (5.3.3) and (5.3.4) we can modify equation (5.3.6) to calculate the required sample size rather than the number of events:

$$N = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{(\log \Delta)^2} \times \frac{(\frac{1}{3})^3}{\pi_1 \pi_2 \pi_3 P_E} \tag{5.3.7}$$

Equation (5.3.7) can be rearranged to isolate for power:

$$1 - \beta = \Phi\left(Z - Z_{\alpha/2}\right), Z = \frac{(\log \Delta)\sqrt{N \pi_1 \pi_2 \pi_3 P_E}}{(\frac{1}{3})^{\frac{3}{2}}} \tag{5.3.8}$$

5.3.3.5 Expected proportions

Under the assumption of Hardy - Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population and the relationship between genotype frequencies and allele frequencies are given by:

$$\begin{aligned} P(AA) &= P(A)P(A) \\ P(Aa) &= 2P(A)P(a) \\ P(aa) &= P(a)P(a) \end{aligned} \tag{5.3.9}$$

Based upon the observed number of patients experiencing survival event during the study for each cohort we calculated that the failure proportion (P_E) for overall survival was 0.31 in the ovarian cohort and 0.64 in the lung cohort. For progression free survival, the failure proportion was 0.70 in the ovarian cohort and 0.82 in the lung cohort.

Assuming a risk allele frequency of 0.2, the expected proportions of each genotype are 0.04, 0.32, 0.64 for the ***aa***, ***Aa*** and ***AA*** genotypes respectively (with ***a*** being the risk allele). Similarly, a risk allele frequency of 0.1 would produce expected genotype proportions of 0.01, 0.18, and 0.81 for

the ***aa***, ***Aa*** and ***AA*** genotypes respectively. Table 5.3.2 presents the hazard ratios that we had 80% power to detect in each cohort for each efficacy endpoint (calculated using equation (5.3.8)):

Table 5.3.2 Minimum HR (conferred by each copy of the risk allele) needed to provide 80% power to detect an association between a SNP the safety phenotype based upon the observed prevalence of cases assuming a frequency of 0.1 or 0.2 for the risk allele (MAF) in the population ($\alpha = 2.1e-7$)

	MAF	***Ovarian***	***Lung***
Overall Survival	0.1	38.70	7.78
	0.2	4.67	2.37
Progression Free Survival	0.1	11.40	6.13
	0.2	2.79	2.15

5.3.4 Efficacy

The most widely used statistical method in the analysis of survival endpoints is the Cox proportional hazards model, which describes the instantaneous risk of failure at time *t* by the hazard rate (see section 3.3.3.1)

The median survival is a commonly used metric to summarise the survival function of a group of patients. A unique feature of survival data is the presence of censored observations due to either the withdrawal of patients or to the termination of the study observation period prior to all patients experiencing the survival event in question (endpoint). The median of such data cannot be analysed by ignoring the censored observations because, amongst other considerations, longer-lived patients are more likely to be censored. Hence ignoring the censored observation would likely result in an underestimate of the true median survival. The Kaplan-Meier method (also called the product-limit method) is used to calculate the survivor function (and confidence intervals) from which the median survival of a group of patients can be estimated. This analysis methodology correctly uses the censored observations in addition to the uncensored observations. The Kaplan-Meier estimate [148] of the survivor function is given by Kalbfleisch and Prentice [237]:

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j} \qquad (5.3.10)$$

Where $t_1 < t_2 < \cdots < t_D$ represent the distinct ordered event times. For each $i = 1, \ldots, D$, n_i is the number of patients remaining in the risk set prior to t_i, and d_i is the number of patients that have experienced the event at t_i.

If we let $s_i = n_i - d_i$, the population of patients for which the event did not occur then the estimate of the standard error corresponding to the survival function is given by:

$$\hat{\sigma}(\hat{S}(t_i)) = \hat{S}(t_i)\sqrt{\sum_{j=1}^{i}\frac{d_j}{n_j s_j}} \tag{5.3.11}$$

The second quartile (median – 50th percentile) of the survival function is the time beyond which 50% of the patients in the population are expected to survive and can be estimated by:

$$q_{.50} = \min\{t_j | \hat{S}(t_j) < 0.50\} \tag{5.3.12}$$

The Kaplan-Meier method was also used to calculate the median follow-up time. By reversing the censored status indicator, i.e. death censors the true, but, unknown, observation time of an individual, and censoring is the endpoint of interest, the product limit estimate now computes the follow-up time [238].

5.3.5 Safety

As described in Chapter 2, all safety endpoints were collected using the Common Terminology Criteria for Adverse Events (CTCAE). CTCAE is a list of terms (adverse events) commonly encountered in oncology interventions. Each AE term is defined and associated with a rating scale of severity that indicates the severity of the AE. While the definition of a dose-limiting toxicity (DLT) is determined by the individual treatment circumstances and not the CTCAE it is typical that Grade 3/4 adverse events based CTCAE represent DLTs even when the specific symptoms can be controlled or ameliorated with appropriate supportive care measures. For this reason, we consider Grade 3 or greater to be 'medically relevant'. Consequently, we recoded each safety endpoint response y_i as binary, assuming only two values that for convenience we code as one or zero.

$$Y_i = \begin{cases} 1 & if\ the\ i^{th}\ individual\ experienced\ a\ grade\ 3\ or\ greater\ event\ at\ any\ cycle \\ 0 & if\ the\ i^{th}\ individual\ experienced\ a\ grade\ 2\ or\ lesser\ event\ at\ any\ cycle \end{cases}$$

The safety events are then modelled using a binary logistic regression model

$$\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\Sigma_{i=1}^{p}\beta_i X_i} \quad (5.3.13)$$

Where:

$\pi_i = P(Y_i = 1|X_i = x_i)$: the probability that the dependent variable equals a case given the value of a predictor.

5.3.6 Two different models

For each SNP, we test that hypothesis that SNP alleles have an effect upon efficacy and/or safety response (but with unknown effect size). Cox proportional hazards models and logistic regression models were fitted to the efficacy and safety endpoints respectively. The term 'effect estimate' therefore refers to hazard ratios in the Cox proportional hazards models and odds ratios in the logistic regression models.

We start out with univariate SNP analyses, i.e. a Cox or logistic regression model including only the individual SNP and top 5 principal components (PCs) as covariates. The inclusion of the principal components will adjust for ethnic differences [239] and allow for the inclusion of all ethnicities without the need to analyse only the Caucasian subset. Hereafter we refer to this analysis as 'SNP Only', despite the inclusion of the PCs in every model:

$$\mathbf{X} = (SNP_i, PC_1, PC_2, PC_3, PC_4, PC_5)^T$$

Next, we performed a set of analyses that were adjusted for age (as a continuous term) and sex in addition to each SNP together with the top 5 PCs – hereafter referred to as 'SNP Age Sex or Full Model':

$$\mathbf{X} = (SNP_i, PC_1, PC_2, PC_3, PC_4, PC_5, Age, Sex)^T$$

For both the efficacy and safety endpoints, the reported *p*-value is from the Wald test [240].

The association results between each SNP and safety / efficacy endpoints are summarised using Manhattan plots. Quantile-quantile (Q-Q) plots were used to examine the observed association against expected association for each analysis.

5.4 Results

5.4.1 Treatment, response and outcomes

Overall, 566 patients (223 ovarian and 343 lung) who received platinum chemotherapy as a first-line therapy were included in the analysis. The demographic and clinical characteristics of patients are described in section 2.4.1. All participants were followed until 05/09/2014. The median follow-up duration (Kaplan-Meier estimates) was 36.76 months (assuming 30.436875 days to a Gregorian month), and 291 patients (51.4%) were deceased by the end of the study. The median OS time was 37.7 months (95% confidence interval (CI), 32.6–43.7 months). In the lung cohort, 221 patients (64.4%) had died by the end of the observation period. In the ovarian cohort, 70 patients (31.4%) had died by the end of the observation period. As patients tend to exhibit cancer progression prior to death, in both cohorts the number of PFS events was greater than the number of OS events (Lung: 275 patients – 81.6%, Ovarian: 148 patients – 69.5%). The number of events and censors by cohort are presented in Table 5.4.1. Table 5.4.2 summarises the survival functions for OS and PFS by cohort.

Figure 5.4.1 shows that the OS for the ovarian patients was significantly longer compared with lung patients. Similarly, Figure 5.4.2 shows that the PFS for the ovarian patients was significantly longer compared with lung patients.

Table 5.4.1 Number and proportion of censors and events by endpoint and cohort

	Overall Survival		***Progression Free Survival***	
	Ovarian	***Lung***	***Ovarian***	***Lung***
Number Events	70 (31.4%)	221 (64.4%)	148 (69.5%)	275 (81.6%)
Number Censored	153 (68.6%)	122 (35.6%)	65 (30.5%)	62 (18.4%)

Table 5.4.2 OS and PFS characteristics of each cohort

Survival Characteristic	***Cohort***	
	Ovarian	***Lung***
Overall Survival, days		
Median (95% CI)	2625 (1811-NE)	788 (685-934)
3rd Quartile (95% CI)	Not reached	1595 (1394-2433)
1-year survival rate (proportion)	0.91	0.72
95% CI	0.87 - 0.94	0.67-0.76
2-year survival rate (proportion)	0.82	0.53
95% CI	0.76-0.86	0.47-0.58
Progression Free Survival, days		
Median (95% CI)	578 (498-709)	299 (267-338)
3rd Quartile (95% CI)	1649 (1063-2172)	641 (530-800)

NE = upper confidence interval of the median survival in the ovarian cohort was not estimable due to the low number of events.

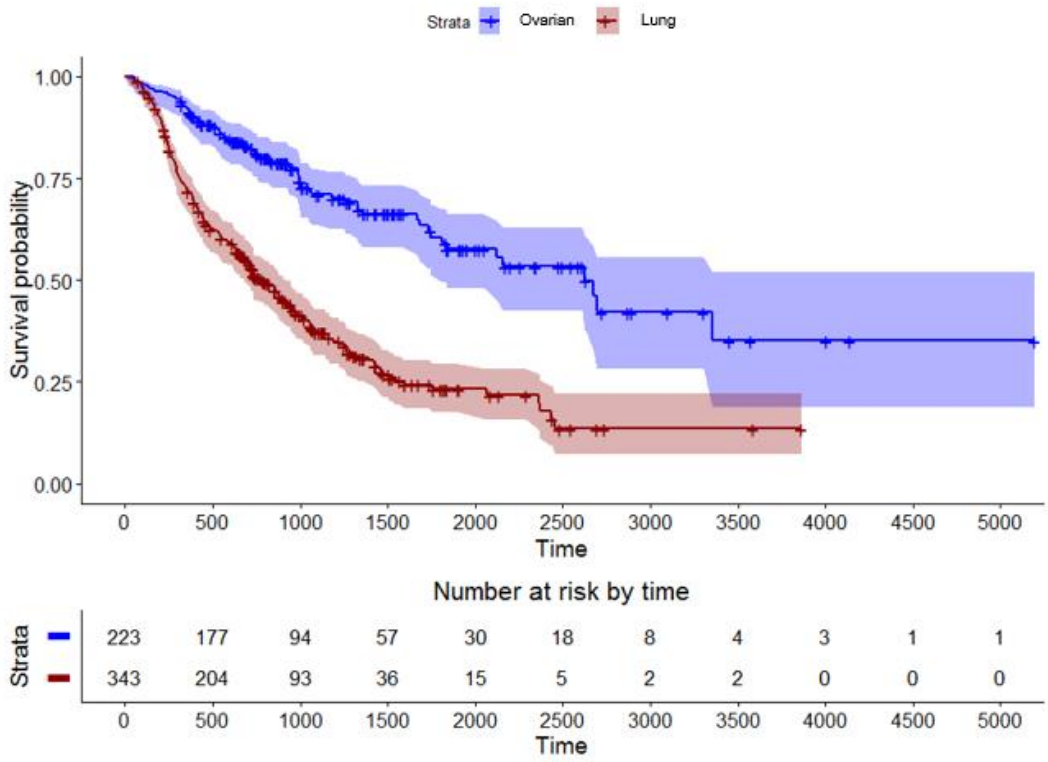


Figure 5.4.1 Kaplan-Meier overall survival curves for the lung and ovarian cohorts

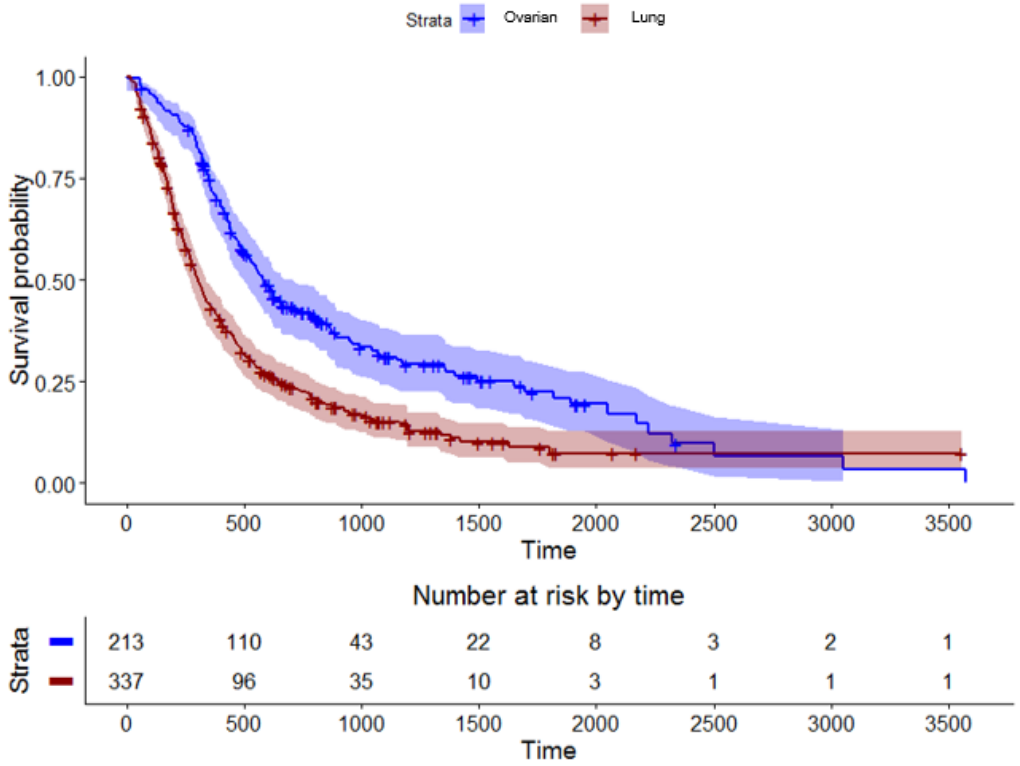


Figure 5.4.2 Kaplan-Meier progression free survival curves for the lung and ovarian cohorts

Table 5.4.3 shows the frequency of neutropenia and gastrointestinal disorder by grade category (0-2, 3-4) by cohort and platinum therapy within cohort. In both lung and ovarian cohorts, grade 3-4 neutropenia (i.e. severe neutropenia) was more common than grade 3-4 gastrointestinal disorder.

Table 5.4.3 Number of patients in each safety endpoint category by cohort

Toxicity and Grade	***Ovarian***	***Lung***		
	Carboplatin	Carboplatin	Cisplatin	All
Neutropenia				
0-2	162	72	151	223
3-4	61	65	55	120
Gastrointestinal Disorder				
0-2	203	121	177	298
3-4	20	16	29	45

5.4.2 GWAS discovery and identification of SNPs associated with survival (efficacy) endpoints

The results of the association between the SNPs and overall survival in the lung cohort are presented in Figure 5.4.3. Panel A presents the 'SNP Only' results, panel B presents the 'full model' results and panel C is an overlay between both plots. Appendix B presents the corresponding figures for OS in the ovarian cohort and PFS, GID and NEU Manhattan plots for both the lung and ovarian cohorts (see Table 5.4.4). Appendix C presents the results for the Q-Q plots matching the data presented in the Manhattan plots. As the points on the lower end of the distribution fall on the reference line, we conclude that the *p*-values follow a normal distribution. This indicates that there was no inflation of test statistics, and that any population stratification was well controlled for using PCs in the analyses.

Table 5.4.4 Manhattan plots contained within Appendix B

Cohort	*Endpoint*	*Figure Number*
Ovarian	Overall Survival	Figure 12.2.1
Ovarian	Progression Free Survival	Figure 12.2.2
Lung	Progression Free Survival	Figure 12.2.3
Ovarian	Neutropenia	Figure 12.2.4
Lung	Neutropenia	Figure 12.2.5
Ovarian	Gastrointestinal Disorder	Figure 12.2.6
Lung	Gastrointestinal Disorder	Figure 12.2.7

Typically, within GWAS studies, only SNPs that are significantly associated (pass multiple testing correction) with the phenotype are presented in the results. As we had so few significant results, and our goal was to compare models that include clinical covariates from those that do not, we present the top 5 most associated results from each phenotype, irrespective of their significance. The 5 SNPs that were most significantly associated with each outcome from each analysis are summarised in Table 5.4.5-Table 5.4.8. Below we detail the results presented in the Figures and Tables.

5.4.2.1 Overall survival

The SNPs that were most significantly associated with OS are summarised in Table 5.4.5. In the ovarian cohort, of the top 5 SNPs from the SNP Only results, 2 SNPs (rs186885699 and rs34422484) remained in the top 5 most associated once age and gender were added as covariates. Two of the top 5 SNPs from each analysis are contained within annotated genes with 3 of the top 5 SNPs from each analysis being intergenic.

Table 5.4.5 Top 5 SNPs from each cohort: Outcome OS

Cohort	RSID	GENE	Chr:BPᵀ	Major/ minor allele	SNP Only†			Full Model†		
					Hazard‡ Ratio	95% CI	*p*-value	Hazard‡ Ratio	95% CI	*p*-value
Ovarian	rs140634372	*PRDM5*	4:121719544	T\A	7.331	3.148-17.07	3.86E-06	.		.
	rs186885699	*MROH5*	8:142480801	G\A	6.652	2.941-15.05	5.34E-06	6.166	2.773-13.71	8.13E-06
	rs34422484		22:50696678	G\A	4.901	2.441-9.840	7.85E-06	5.321	2.641-10.72	2.91E-06
	rs2234455		11:5529152	A\G	2.465	1.613-3.767	3.07E-05	.		.
	rs73974725		17:10409241	A\G	5.706	2.508-12.98	3.30E-05	.		.
	exm2271283*		10:47646751	G\A	.		.	2.256	1.508-3.376	7.59E-05
	rs140112498		22:38026059	G\A	.		.	6.511	2.656-15.96	4.21E-05
	rs16933415	*CPA6*	8:68497939	A\G	.		.	3.170	1.803-5.572	6.13E-05
Lung	rs17117678	*OMA1*	1:58999651	A\C	2.431	1.789-3.304	1.36E-08	2.338	1.714-3.188	8.11E-08
	rs35075952	*TACSTD2*	1:59042311	A\C	2.402	1.764-3.269	2.57E-08	2.304	1.686-3.148	1.63E-07
	rs2001030		26:1438	G\A	2.496	1.691-3.687	4.23E-06	2.395	1.617-3.547	1.32E-05
	rs5741809		20:36956026	A\G	3.613	1.950-6.695	4.46E-05	3.565	1.974-6.439	2.51E-05
	rs2076212		22:44322970	C\A	1.674	1.294-2.167	8.92E-05	.		.
	rs17130717	*GBP1*	1:89524657	G\C	.		.	2.560	1.589-4.125	1.13E-04

T: Chr indicates the chromosome number and BP indicates position in base pairs

†: All analyses included the top five principal components as covariates

‡: The coding of the Cox analysis was such that the hazard ratio represents the increase in hazard for each copy of the minor allele

*: I used the Illumina RSID mapping spreadsheet to convert each Illimuna exm-ID back to an RSID. Not all tags had an equivalent RSID, exm2271283 was one such SNP

The results of the association between the SNPs and OS for the lung cohort are presented in Figure 5.4.3. In the lung cohort, of the top 5 SNPs from the SNP Only results, 4 SNPs (rs17117678, rs2001030, rs35075952 and rs5741809) remained in the top 5 most associated once age and sex were added as covariates. Similar to the ovarian cohort, 2 of the top 5 SNPs from each analysis are located within annotated genes while the remainder are intergenic.

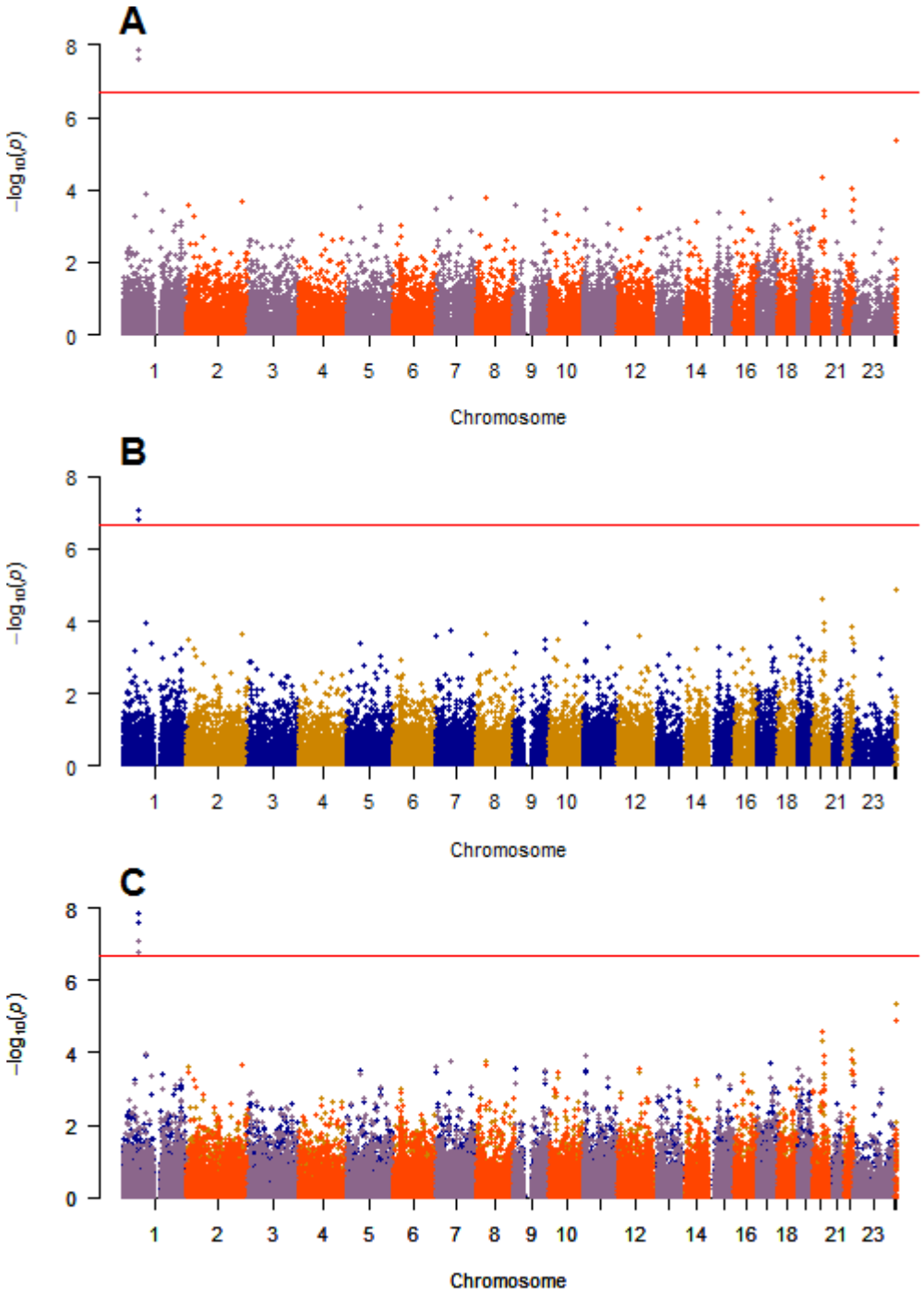


Figure 5.4.3 Exome array analysis results of overall survival in the lung cohort.
The association of SNP genotype and overall survival was evaluated using a Cox regression models for 29328 SNPs across 343 patients with lung cancer treated with platinum-therapy containing regimens. regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B). (P = 2.1 × 10−7; red line).

Figure 5.4.4 presents the results for the top 5 SNPs from each analysis. In the majority of SNPs that overlap between analyses, the addition of covariates does not improve the association signal strength.

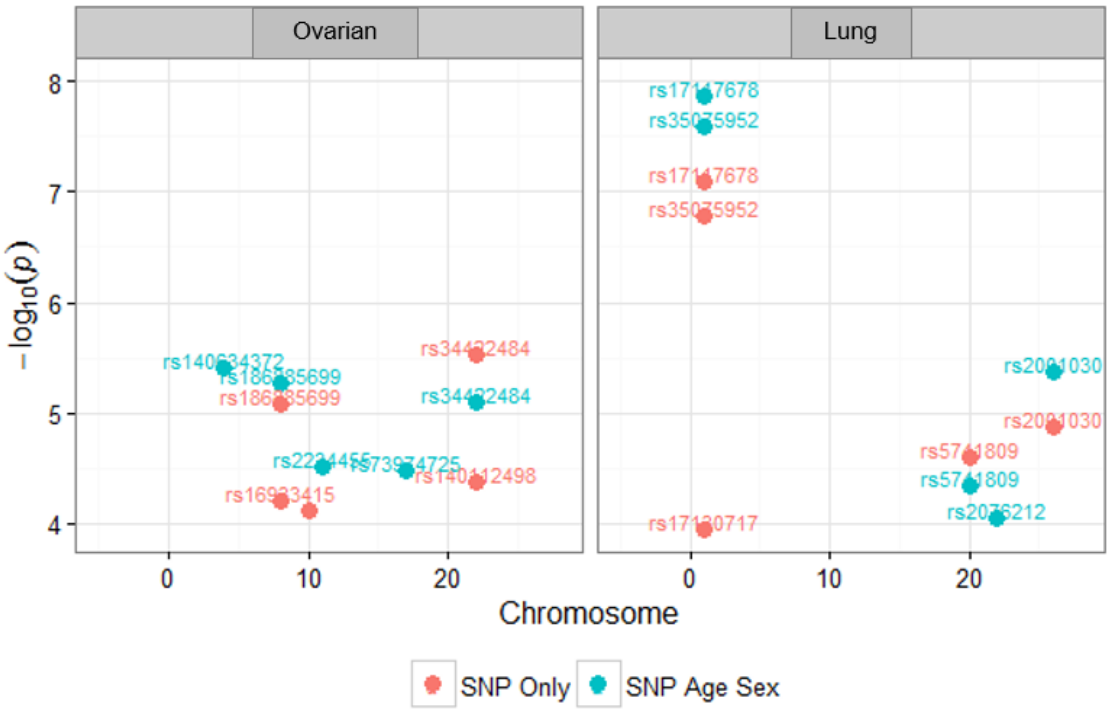
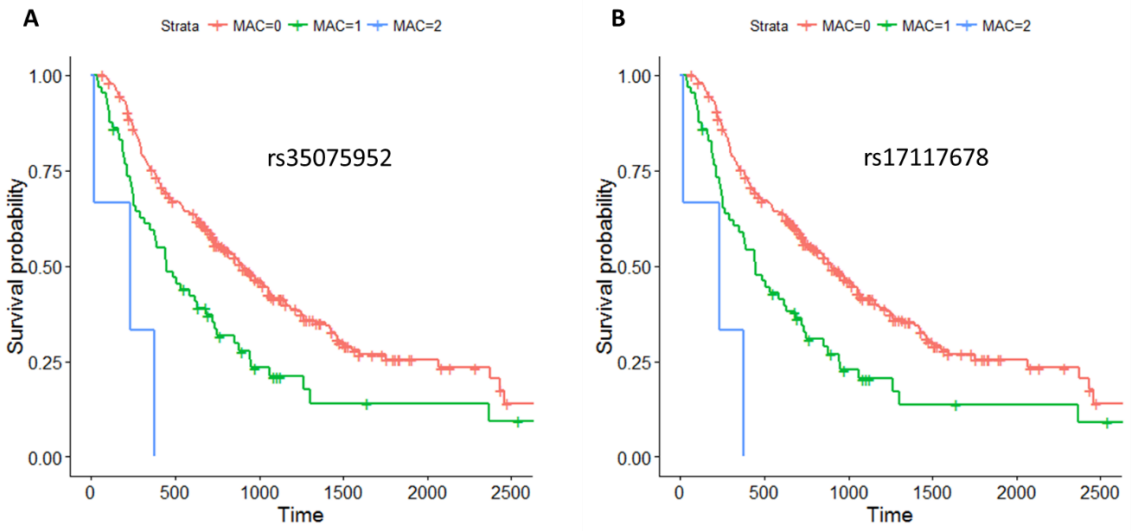


Figure 5.4.4 Five most OS associated SNPs by cohort.
SNP Only: SNP + top 5 principal components; SNP Age Sex: SNP + top 5 principal components + Age + Sex.

Two SNPs reached genome wide significance in the lung cohort SNP Only analysis (rs1711767, *p*=1.36E-08; and rs35075952; *p*=2.57E-08). Both SNPs remained significant at the predefined alpha threshold (2.1e-7) with the addition of clinical covariates (rs1711767, *p*=8.11E-08; and rs35075952; *p*=1.63E-07). We tested each SNP for proportional hazard by including a minor allele count × time interaction [241], neither SNP failed the proportional hazard assumption in the SNP only model (p=0.0844 and 0.0807 for the rs35075952 and rs1711767 SNPs respectively) or with the addition of age and sex covariates (p=0.3647 and 0.3551 for the rs35075952 and rs1711767 SNPs respectively). The Kaplan–Meier plots of these two SNPs stratified by minor allele count (MAC) are presented in Figure 5.4.5. The most significant association was with rs1711767, located in the *OMA1* gene. A regional plot of rs1711767 reveals that it is in strong linkage disequilibrium (LD) with rs35075952 (r^2=0.97), so these are not independent signals of association (Figure 5.4.6).



MAC	N	Events	Median	95% LCI	95% UCI
0	275	167	901	734	1056
1	65	51	449	289	686
2	3	3	234	20	NA

MAC	N	Events	Median	95% LCI	95% UCI
0	276	167	901	734	1056
1	64	51	445	289	634
2	3	3	234	20	NA

Figure 5.4.5 Kaplan-Meier plots of overall survival in the lung cohort for top two associated SNPs stratified by minor allele count.
Plots and summary tables show the relationship between genotype at (A) *TACSTD2* rs35075952; (B) *OMA1* rs17117678 and overall survival in the lung cohort. MAC refers to the number of copies of the risk/minor allele. For both SNPs, variant carriers had reduced survival.

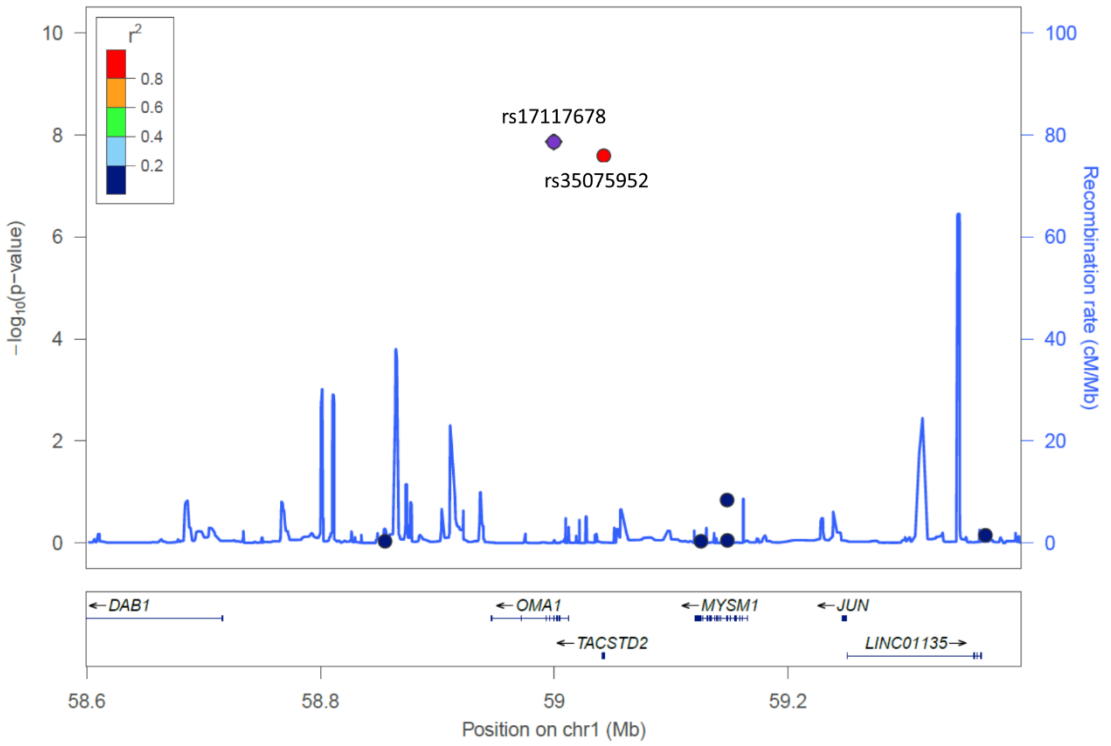


Figure 5.4.6 SNP-OS association results in the lung cohort.
The plot shows association results for SNPs ($-\log_{10}$ *p*-value) as a function of genomic distance. The purple diamonds indicates the SNP with the strongest association evidence (rs17117678). Each circle represents a SNP, with the colour of the circle indicating the correlation between that SNP and the index SNP (purple diamond). These results suggest rs35075952 is in strong LD with the index SNP (rs17117678) and therefore the association that each SNP has with overall survival is likely not independent.

5.4.2.2 Progression free survival

In the ovarian cohort, of the top 5 SNPs from the SNP Only analysis, 3 SNPs (rs3751057; rs62007840 and rs7212197) remained in the top 5 most associated once age and gender were added as covariates. None of the most associated SNPs from either analysis was contained within annotated genes (see Table 5.4.6) and none reached our Bonferroni-corrected threshold of significance.

In the lung cohort, of the top 5 SNPs from the SNP Only analysis, 2 SNPs (rs1478091 and rs34379253) remained in the top 5 most associated once age and sex were added as covariates. Similar to the ovarian cohort, 1 SNP from each analysis are located within annotated genes while the remainder are intergenic.

Figure 5.4.7 presents the results for the top 5 SNPs from each analysis. As with OS, where the SNPs are within the top 5 associated for both SNP Only and SNP Age Sex analyses, the addition of clinical covariates does not improve the association strength signal.

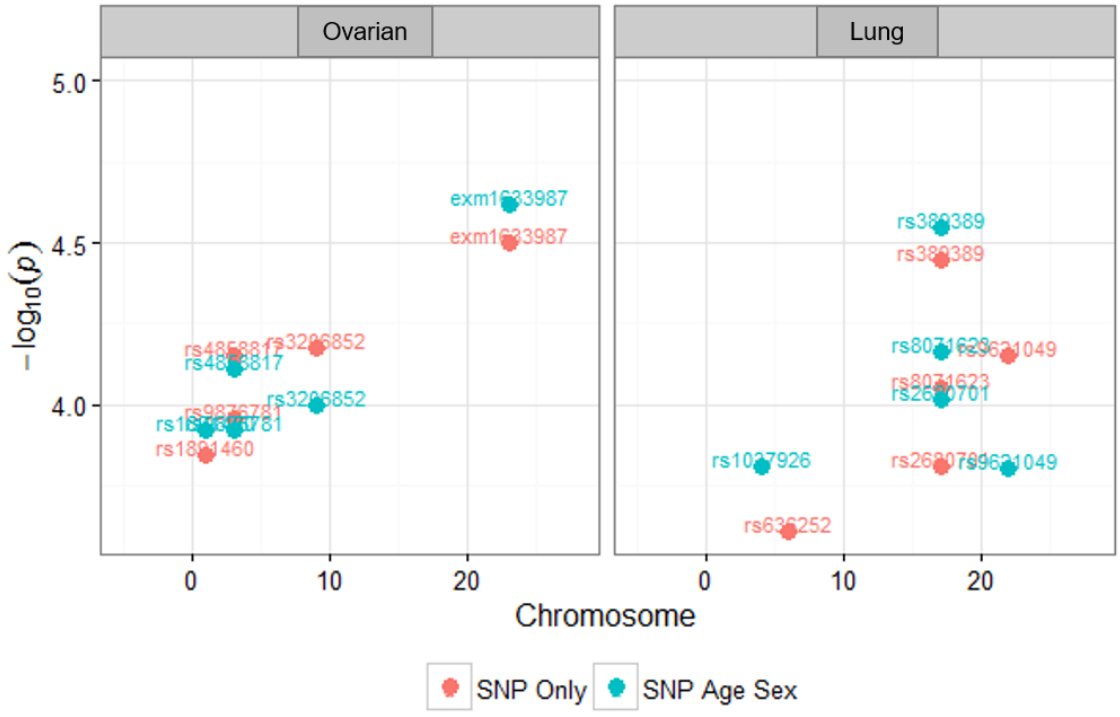


Figure 5.4.7 Five most PFS associated SNPs by cohort.
SNP Only: SNP + top 5 principal components; SNP Age Sex: SNP + top 5 principal components + Age + Sex.

5.4.2.3 Neutropenia

In the ovarian cohort, all the top 5 SNPs from the SNP Only analysis remained in the top 5 most associated once age and gender were added as covariates. Three of these SNPs were contained within annotated genes (see Table 5.4.7).

In the lung cohort, of the top 5 SNPs from the SNP Only analysis, 4 SNPs remained in the top 5 most associated once age and sex were added as covariates. Only 1 SNP from the SNP Only analysis is located within an annotated gene.

Figure 5.4.7 presents the results for the top 5 SNPs from each analysis. The addition of clinical covariates does not alter the association strength of the SNPs with neutropenia.

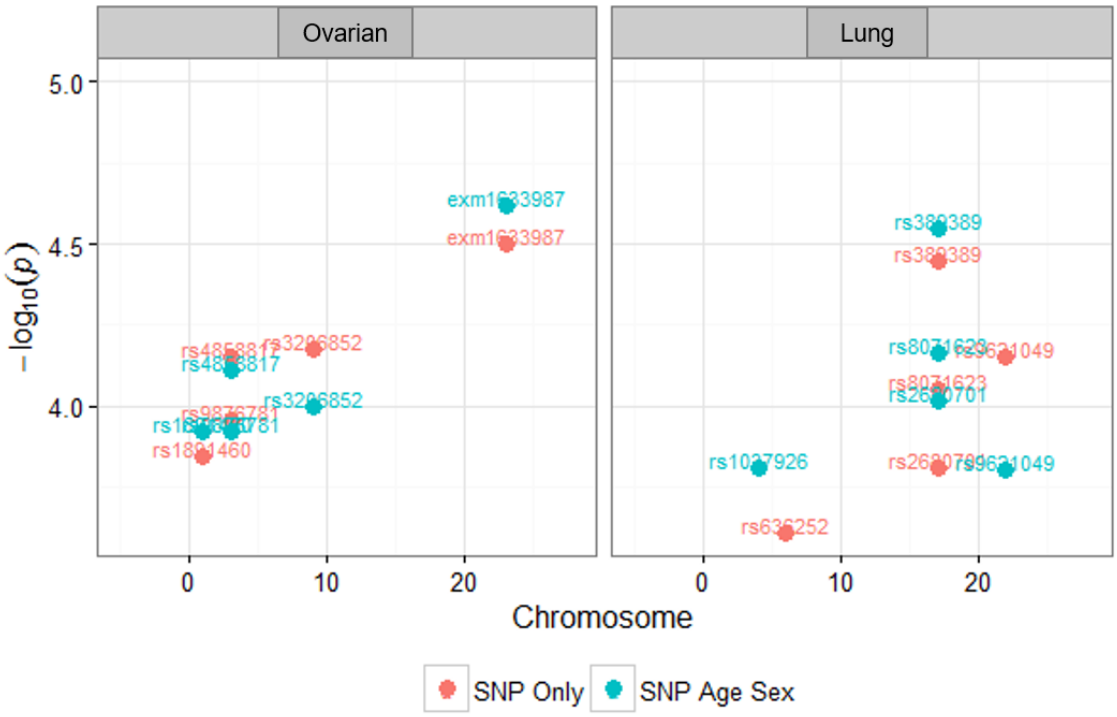


Figure 5.4.8 Five most neutropenia associated SNPs by cohort.
SNP Only: SNP + top 5 principal components; SNP Age Sex: SNP + top 5 principal components + Age + Sex.

5.4.2.4 Gastrointestinal disorders

In the both the ovarian cohort and lung cohort, all of the top 5 SNPs from the SNP Only analysis remained in the top 5 most associated once age and gender were added as covariates. Three of these SNPs were contained within annotated genes (see Table 5.4.7).

Figure 5.4.7 presents the results for the top 5 SNPs from each analysis. The addition of clinical covariates does not alter the association strength of the SNPs with neutropenia.

Figure 5.4.9 Five most gastrointestinal disorder associated SNPs by cohort.
SNP Only: SNP + top 5 principal components; SNP Age Sex: SNP + top 5 principal components + Age + Sex.

Table 5.4.6 Top 5 SNPs from each cohort: outcome PFS

Cohort	RSID	GENE	Chr:BPT	Major/ minor allele	SNP Only†			Full Model†		
					Hazard‡ Ratio	95% CI	*p*-value	Hazard Ratio	95% CI	*p*-value
Ovarian	rs7212197		17:48335925	A\G	0.534	0.325-0.723	1.65E-05	0.507	0.316-0.686	3.69E-06
	rs62007840		15:78570918	A\G	0.593	0.471-0.924	3.66E-05	0.613	0.496-0.974	1.32E-04
	rs2491014		23:138897130	C\A	4.246	0.719-5.794	5.40E-05	.		.
	rs13057352		22:47095235	C\A	3.863	0.689-4.600	5.57E-05	.		.
	rs3751057		11:9051475	C\G	2.615	0.715-2.738	7.41E-05	2.761	0.797-3.082	3.64E-05
	exm2268202*		19:48800338	G\A	.		.	1.842	1.443-3.209	2.23E-05
	rs12937557		17:77078069	G\A	.		.	4.141	0.863-4.770	7.03E-06
Lung	rs1478091		4:131790501	A\G	2.117	1.205-2.369	4.64E-06	2.135	1.173-2.306	3.89E-06
	rs34379253		16:57449687	G\A	2.508	1.003-2.565	2.02E-05	2.471	0.941-2.421	2.91E-05
	rs35075952	*TACSTD2*	1:59042311	A\C	1.810	1.764-3.269	4.77E-05	.		.
	rs74991234		18:52604083	A\G	2.166	1.315-2.880	5.69E-05	.		.
	rs9546785		13:36801415	G\A	0.693	0.622-0.923	6.51E-05	.		.
	rs2076212		22:44322970	C\A	.		.	1.688	1.276-2.144	6.23E-05
	rs2286455	*PROM1*	4:16020162	G\A	.		.	1.853	0.984-1.855	3.88E-05
	rs5741809		20:36956026	A\G	.		.	2.916	1.974-6.439	5.26E-05

T: Chr indicates the chromosome number and BP indicates position in base pairs

†: All analyses included the top five principal components as covariates

‡: the coding of the Cox analysis was such that the hazard ratio represents the increase in hazard for each copy of the minor allele

*: I used the Illumina RSID mapping spreadsheet to convert each Illimuna exm-ID back to an RSID. Not all tags had an equivalent RSID, exm2271283 was one such SNP

Table 5.4.7 Top 5 SNPs from each cohort: outcome neutropenia

Cohort	RSID	GENE	Chr:BP[T]	Major/ minor allele	SNP Only[†]			Full Model[†]		
					Odds Ratio	95% CI	*p*-value	Odds Ratio	95% CI	*p*-value
Ovarian	exm1633987 *		23:36371719	A\T	5.246	2.430-11.32	2.42E-05	5.123	2.374-11.06	3.15E-05
	rs4858817	*FBXW12*	3:48416756	G\A	2.692	1.647-4.399	7.78E-05	2.719	1.660-4.454	7.07E-05
	rs3206852	*FOCAD*	9:20953049	A\G	2.650	1.622-4.331	1.00E-04	2.767	1.678-4.564	6.67E-05
	rs9876781		3:48487338	G\A	2.567	1.588-4.150	1.19E-04	2.594	1.600-4.205	1.10E-04
	rs1891460	*PM20D1*	1:205814497	G\A	4.490	2.089-9.652	1.20E-04	4.450	2.062-9.603	1.42E-04
Lung	rs389389		17:56676368	A\G	2.451	1.611-3.730	2.84E-05	2.445	1.600-3.736	3.57E-05
	rs8071623		17:56621286	A\C	2.294	1.524-3.452	6.87E-05	2.284	1.511-3.451	8.85E-05
	rs2680701		17:56438301	G\A	2.410	1.549-3.750	9.64E-05	2.362	1.513-3.687	1.55E-04
	rs1037926	*CCSER1*	4:91227681	G\A	0.515	0.365-0.726	1.55E-04	.		.
	rs9621049		22:31013419	G\A	2.508	1.557-4.040	1.58E-04	2.680	1.649-4.358	7.02E-05
	rs636252		6:117157774	G\A	.		.	1.862	1.336-2.595	2.45E-04

T: Chr indicates the chromosome number and BP indicates position in base pairs

†: All analyses included the top five principal components as covariates

*: I used the Illumina RSID mapping spreadsheet to convert each Illimuna exm-ID back to an RSID. Not all tags had an equivalent RSID, exm2271283 was one such SNP

Table 5.4.8 Top 5 SNPs from each cohort: outcome gastrointestinal disorder

Cohort	RSID	GENE	Chr:BP[T]	Major/ minor allele	SNP Only[†]			Full Model[†]		
					Odds Ratio	95% CI	*p*-value	Odds Ratio	95% CI	*p*-value
Ovarian	rs78367010		10:135195148	G\A	21.026	5.002-88.39	3.22E-05	20.499	4.816-87.25	4.36E-05
	rs10889205	*RLF*	1:40705726	A\T	5.204	2.369-11.43	4.01E-05	6.193	2.708-14.16	1.56E-05
	rs35032920	*OSBPL6*	2:179188959	G\A	10.125	3.291-31.16	5.42E-05	11.066	3.521-34.78	3.88E-05
	rs2781806		10:116060427	C\G	16.455	3.763-71.96	1.99E-04	19.436	4.226-89.39	1.38E-04
	rs1891110		10:124610027	A\G	0.198	0.079-0.497	5.62E-04	0.191	0.075-0.485	4.91E-04
Lung	rs3815045		11:61669946	G\A	11.858	3.540-39.72	6.08E-05	13.620	3.911-47.43	4.09E-05
	rs900399		3:156798732	A\G	2.685	1.641-4.393	8.49E-05	2.739	1.667-4.499	6.97E-05
	rs900400		3:156798775	A\G	2.685	1.641-4.393	8.49E-05	2.739	1.667-4.499	6.97E-05
	rs1539172	*CCDC171*	9:15784631	G\A	2.619	1.603-4.280	1.22E-04	2.875	1.710-4.834	6.82E-05
	rs11806429	*PLEKHG5*	1:6550505	G\A	3.448	1.752-6.789	3.41E-04	3.765	1.875-7.561	1.94E-04

T: Chr indicates the chromosome number and BP indicates position in base pairs

†: All analyses included the top five principal components as covariates

5.5 Discussion

The motivation behind this study was to examine if the addition of baseline covariates could improve our ability to detect SNP associations with efficacy and/or safety outcomes. The approaches and scenarios examined represent standard analysis techniques and typical data for which the decisions of covariate inclusion are made on a regular basis. Our analyses indicate that the introduction of covariates to the variable of interest (SNP) does not improve the power to detect SNP-phenotype interactions.

The purpose of adjustment for covariates in genome-wide association studies (GWAS) is to account for potential confounding factors that can bias SNP effect estimates, and/or to improve statistical power by reducing residual variance [208, 242]. Within genetic association studies, it is standard to adjust for ancestry-informative principal components calculated from individual genotypes in order to account for potential confounding from population structure [208, 242]. Failure to adjust for confounding factors can result in spurious associations [208]. Besides confounding factors, it is common practice in medical research to adjust for demographic factors such as sex and age in an effort to increase statistical power [243, 244]. We contend that the rationale for the inclusion of these demographic factors is the commonly held belief that if they truly are predictive of the outcome then they should decrease the residual variance of the outcome variable and therefore increase the magnitude of the SNP effect size relative to the phenotypic variance which will lead to an increase in power [231].

In a linear regression model (as would be appropriate for a continuous phenotype), inclusion of an independent (of genotype) predictive covariate would not change the magnitude of the estimated SNP effect. It would however improve the precision of the estimator thereby increasing statistical power [233]. In contrast, the effect of covariates is not so straightforward in the context of non-linear models like logistic and Cox regression methods.

Logistic regression differs from linear regression in that inclusion of an independent predictive covariate both increases the magnitude of the estimated effect and simultaneously reduces the precision of the effect estimator [207, 245]. The consequence for power depends on the relative increase in the estimated effect (which would increase power) against the inflation of the estimator variance (decrease in precision - which would reduce power) [246]. Similarly, for Cox proportional

hazards models, the inclusion of a predictive covariate will not always increase power [247]. This loss of power is attributed to either non-independence between the SNP effect and the other covariates in the model [247]; or the fact the covariate itself might not have a consistent multiplicative effect upon hazard across the range of the covariate. Hsieh et al. [233, 246, 247] noted that as the use of covariates increases, the variance of the estimated effect size for the parameter of interest also increased, and thus proposed increasing the sample size for a PH model by a variance inflation factor when covariates are included in the model, even when the covariates are necessary for the model assumptions to hold.

In summary, use of covariates in the analyses resulted in smaller SNP-phenotype association signals in both of the non-linear analytic approaches used, although the impact of their use was generally modest. The use of covariates in the analysis of medical studies often is guided by our understanding of linear models where the inclusion of prognostic factors increases power to detect an association between the response and the primary independent variable. In studies employing non-linear methods, the inclusion of covariates can reduce power to detect an association between the response and the primary independent variable. Consequently, only confounding factors needed to produce an unbiased estimate of effect should be included in the model as covariates. For models employing Cox regression, prognostic factors may help with achieving proportional hazards in which case their inclusion is necessary to meet the model assumptions however researchers should be aware of the implications on power that are created in these circumstances.

We identified two SNPs within the *OMA1* and *TACSTD2* genes that were significantly associated with overall survival in lung cancer patients receiving platinum-based chemotherapy.

The gene *OMA1* encodes a metalloprotease, an enzyme which plays a role in regulation of the inner membrane of mitochondria. Several authors had already identified *OMA1* as a colorectal cancer driver gene – a gene in which genetic changes increase the mutation rate in the cell leading to more rapid tumour evolution and metastases [248]. However, these studies relied upon tumour tissue where our genotyping methods would only capture germline variation. Kong et al. [249] examined cisplatin sensitive and cisplatin resistance ovarian and cervical cell lines. The authors found that cisplatin exposure induces L-Opa1 processing and mitochondrial fragmentation in chemo-sensitive but not in chemo-resistant cells. As Metallopeptidase *OMA1* is

involved in L-Opa1 processing, the authors implicate *OMA1* as playing a role in cisplatin resistance. Within our data, none of the ovarian cancer patients was receiving cisplatin and the majority of patients in the lung cohort were receiving cisplatin. If the effect of *OMA1* were limited to cisplatin and not the other platinum agents, then this might explain the discrepancy between the lung and ovarian association results. Alternatively, the median survival time is much greater for ovarian cancer as compared with lung cancer. This results in fewer survival events for ovarian cancer and therefore lower power to detect SNP-OS associations. The effect sizes for these two SNPs were in the opposing directions for each cohort. As presented in Table 5.4.5, for SNP rs17117678, in the lung cohort the hazard ratio was 2.431 and 2.338 for the SNP Only and full models respectively. In the ovarian cohort, the corresponding hazard ratios were 0.581 and 0.589 (both were non-significant, p=0.138 and p=0.153). Similarly, for SNP rs35075952, in the lung cohort the hazard ratio associated with variant carriers in the lung cohort was 2.402 and 2.304 for the SNP Only and full models respectively. In the ovarian cohort, the corresponding hazard ratios were 0.567 and 0.570 (both were non-significant, p=0.122 and p=0.132).

TACSTD2 encodes the protein Tro*p*-2 which is a carcinoma-associated antigen. The antigen is a cell surface receptor that transduces calcium signals. It is universally expressed in stratified squamous epithelia of many organs, including skin, oesophagus and cervix. It is known as a carcinoma-associated antigen as it is frequently over-expressed in epithelial tumours. For this reason, it is being considered as a molecular target for cancer therapy. Similar to *OMA1*, *TACSTD2* is theorised to play a role in cell apoptosis, and its over-expression is associated with chemo-resistance [250].

Our regional plot shows that the *OMA1* and *TACSTD2* SNPs are in strong linkage disequilibrium. Unfortunately, the exome chip used for genotyping has low coverage of SNPs and none of the other tagged SNPs was in strong enough LD to observe supporting association signals for either significant SNP. The fact that the only other SNP in high LD with the top associated SNP also exhibited a similar statistical association with the phenotype supports the idea that one or both signals are possibly valid but without replication it remains a distinct possibility that one or both are chance findings.

Given the greater number of PFS events as compared with OS events was surprising that we were able to detect OS-SNP associations but not PFS-SNP associations. We attribute this to

error in the assessment of progression. Unlike death (an unambiguous endpoint), there are several potential sources of measurement bias or variability common to the assessment of progression [251]. The exact date of progression cannot be known, since it is determined based on the types and timing of assessments. At the point in time at which a patient is identified as being progressed, it is only known that the progression occurred at some point in the interval of time between the last negative progression assessment and the assessment at which progression was detected. Within the patient records, progression date is recorded as the date of the evaluation at which progression was first evident. While this is consistently an overestimate of PFS, the degree to which it is an overestimate is subject to patient variability in the frequency and timings of progression assessments. Stochastic asymmetry between each genotype with respect to the frequency of assessment has the potential to introduce bias into SNP-PFS associations. Lastly, PFS is a composite endpoint including radiologic assessment, death, or symptomatic progression. Consequently, 'progression' is an event prone to reading error by the interpreter of the clinical assessment items [251].

Chapter 6. GWAS concordance

> Numerical quantities focus on expected values, graphical summaries on unexpected values.
>
> – John Tukey

6.1 Abstract

Background

Complex diseases give rise to multiple and overlapping phenotypes. It is common to analyse each phenotype separately which gives rise to multiple sets of SNP-phenotype association results. This chapter develops a novel method to compare association signals from univariate SNP analyses to understand the concordance/discordance in signal strength and pattern of SNPs for association in two phenotypes.

Methods

The methods are tested on simulated data and an analysis of a previously conducted study which aimed to identify variants associated with toxicity response and survival in patients treated with platinum therapy for either lung or ovarian cancer.

Results

Assessing joint distribution of *p*-values from two association studies across a range of critical thresholds gives rise to a characteristic 'moth plot'. This plot comprises 4 curves – 2 power curves that describe the proportion of concordance between phenotypes and 2 quadratic curves that characterise the discordance in association results between phenotypes. The difference in maximum amplitude between the quadratic discordance curves describes the difference in the mean *p*-value between the two phenotypes. Both the difference between the intercept of the concordance power functions and the maximal height of the discordance functions describe the correlation between the association signals of the two datasets.

Conclusion

This method provides a qualitative assessment method to graphically explore both the correlation and difference in association result distributions when comparing the results of GWAS analyses on different phenotypes.

6.2 Introduction

Genome-wide association studies (GWAS) have emerged as popular tools for identifying genetic variants that are associated with disease risk, quantitative phenotypes, and treatment response. Standard analysis of a case-control GWAS involves assessing the association between each genotyped SNP and phenotype. However, it is clear that substantial pleiotropy exists in the human genome, with many SNPs contributing to multiple phenotypes [252]. Within the field of oncology, survival has traditionally been considered the gold-standard endpoint for cancer clinical trials [253]. Consequently, a large number of cancer GWAS studies have focused on the relationship between single nucleotide polymorphisms (SNPs) and the survival endpoints: overall survival (OS) [254-257] and progression free survival (PFS) [258-260]. More recently there has been a shift towards utilisation of GWAS studies to identify novel common genetic variations for drug-induced toxicity [251, 260-265]. Both survival and toxicity represent clinically relevant cancer outcomes, yet the link between each type of endpoint is poorly understood. Recent GWAS advances have made it possible to jointly model the association between SNPs and several phenotypes [266-268]. While these methods are more powerful than univariate testing methods for detecting significant associations between SNPs and phenotypes, they do not characterise the genome-wide similarity of genetic architecture between phenotypes.

We introduce a new method that compares the results of pair-wise univariate GWAS by exploring the concordance and discordance of association signals between two phenotypes. This is achieved by assessing the proportion of concordant and discordant association results at a series of *p*-value thresholds. The goal of this method is to provide a graphical tool that allows a researcher to simultaneously assess the similarity in the distribution of association tests between two phenotypes and to provide a measure of the correlation of *p*-values between phenotypes even when the distributions differ.

6.3 Methods

6.3.1 Concordance / discordance

We assume an association study has been performed for two endpoints (phenotype 1, phenotype 2), with *p*-values SNPs P_{E1} and P_{E2} for N SNPs. To assess the similarity of association between

SNPs and two alternative endpoints, we define a 2×2 contingency table of possible outcomes for observations between the phenotypes at a fixed significance threshold (α). These outcomes are provided in Table 6.3.1 give the proportion of SNPs classified into each cell, where *A* gives the number of SNPs where both Pp_{E1}≤ α and $Pp_{E2} \leq$ α, *B* gives the number of SNPs where $Pp_{E1} \leq$ α and Pp_{E1}> α, and similarly for *C* and *D*.

Table 6.3.1 Proportion of SNPs categories by α significance threshold for the association results between two phenotypes

		Phenotype 2	
		$p \leq \alpha$	$p > \alpha$
Phenotype 1	$p \leq \alpha$	A	B
	$p > \alpha$	C	D

The table offers a snapshot of all SNPs at a single *α*-threshold. We seek to explore how the relative proportions of each cell change across *α*-thresholds.

A key feature of this method is understanding how the frequency/proportion of each cell of the contingency table is **expected** to change by α, under different assumptions of the *p*-value distribution for each phenotype, and the correlation between them. Assuming that the GWAS *p*-values from each phenotype are approximately uniformly distributed, then at a low α-threshold very few SNPs will fall into cell '*A*' category and the majority of SNPs will have *p*>α for both phenotypes and therefore in the '*D*' category. As the threshold increases the proportion of SNPs in *A*, *B* and *C* is expected to increase until a threshold of *p*=0.5 when there should be equal number of SNPs in each category (*A, B, C* and *D*) (see Figure 6.3.1).

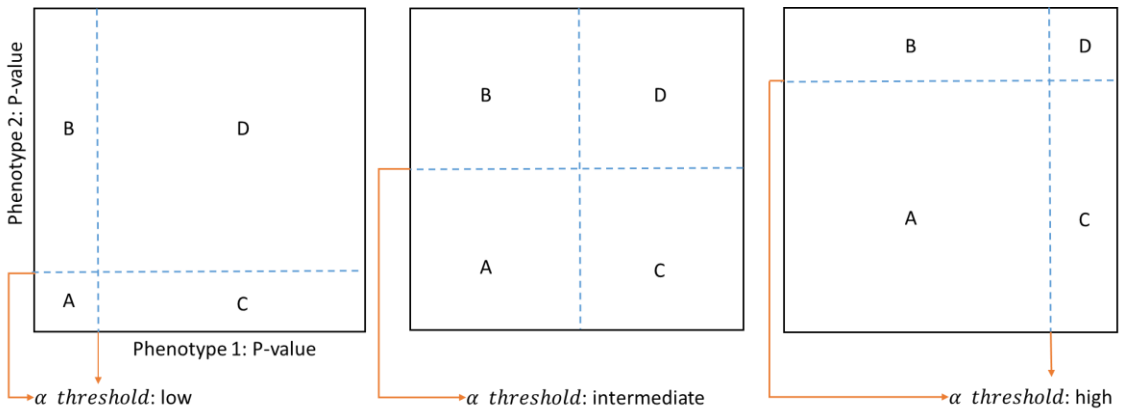


Figure 6.3.1 Schematic for SNP classification by phenotype, where alpha threshold controls the number of SNPs that fall within each category.
This figure is illustrative and makes no assumptions about the underlying distribution of *p*-values. The proportion SNPs within each category at a given threshold will depend on the specific distributions of the *p*-values.

The relative proportion of each cell can be calculated by dividing each cell by the total number of examined SNPs (*A*+*B*+*C*+*D*). When the α-threshold is close to zero, no SNP has a *p*-value from either phenotype that is below the threshold and all SNPs lie in the *D* category. When the α-threshold is close to one, all of the SNPs have a *p*-value less than α in both phenotypes and the proportion of SNPs in category *A* is one.

Under the null hypothesis of no association at any SNP for either phenotype, and no correlation between the test statistics for phenotypes 1 and 2, the relationships between the proportions of SNPs in $[A \quad B \quad C \quad D]$ can be easily constructed.

Assuming *p*-values have a uniform U(0,1) distribution, then:

$P(P_{Ei} \leq \alpha) = \alpha$, for $i = 1, 2$.

Given P_{E1} and P_{E2} are independent,

$A = P(P_{E1} \leq \alpha, P_{E2} \leq \alpha) = P(P_{E1} \leq \alpha)\, P(P_{E2} \leq \alpha) = \alpha^2$

$B = P(P_{E1} \leq \alpha, P_{E2} > \alpha) = P(P_{E1} \leq \alpha)\, P(P_{E1} > \alpha) = \alpha(1-\alpha)$

And similarly, $C=B$, $D=1-A$.

Figure 6.3.2 plots the relative proportion of each category against *α*-threshold. Hereafter we refer to this type of plot as a 'moth' plot.

We hypothesise that both the correlation between *p*-values from each phenotype and distributional differences between the *p*-values from each phenotype will alter the shape and intercept of the contingency cell functions in the moth plot. To explore this hypothesis, we create simulated *p*-value results that are detailed in the next section.

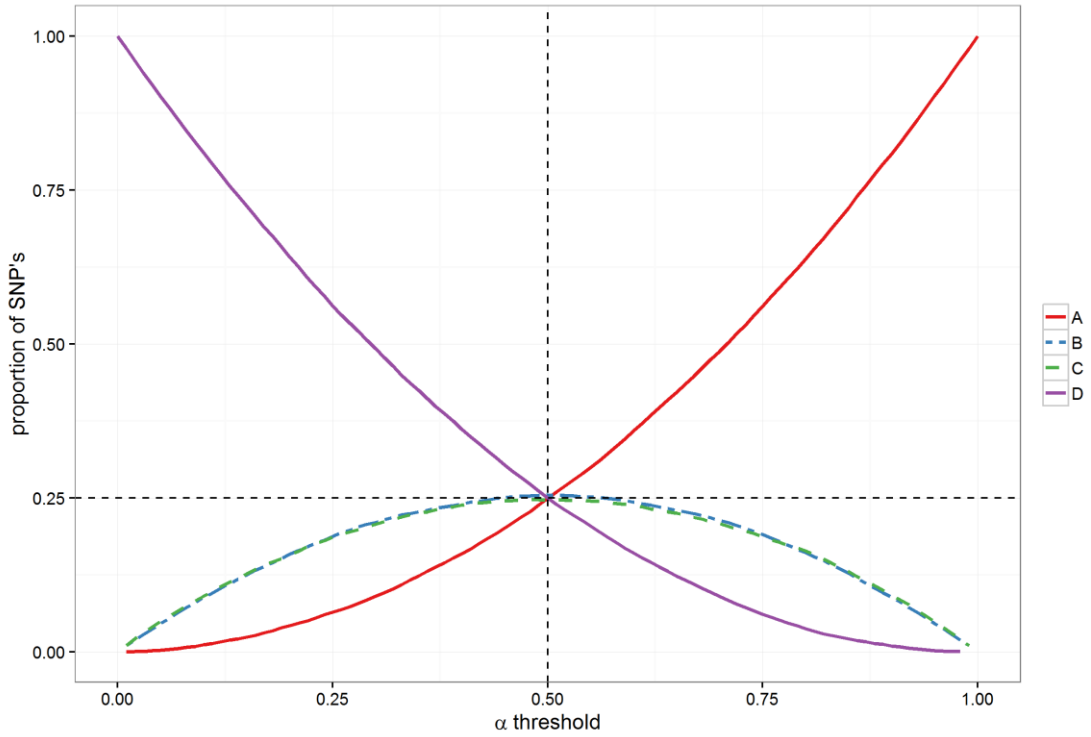


Figure 6.3.2 Moth plot showing the relative proportion of SNPs in each contingency table across a varying α threshold.
A increases as the α threshold increases, and *D* decreases. *B* & *C* both increase from zero until an alpha threshold of 0.5 after which the curves decline. All lines intercept at (0.5,0.25).

6.3.2 Simulation

To explore the behaviour of the moth plot further, we performed a simulation study for test statistics in two phenotypes, varying the level of association within a phenotype and the correlation between phenotypes. We simulated data at 3 levels of correlation (0, 0.5 and 0.9) between *p*-values. Each scenario was simulated for 30,000 SNPs under both a null (H_0) and alternative hypotheses (H_A). Polygenic phenotypes are characterised by modest allele frequencies shifts at many loci [269]. The consequence of this is a shift in the distribution of association *p*-values away from a uniform distribution with more SNPs having lower *p*-values; we refer to this phenomenon as 'polygenic component'; that is to say that there are multiple associations of varying strengths between SNPs and the phenotype being studied. In comparing the pattern of association between SNPs across two phenotypes, the H_0 becomes: there is no

difference in the distribution of *p*-values (polygenic component) between Phenotype 1 and Phenotype 2; while H_A can be written as: there is a difference in the distributions between *p*-values for Phenotype 1 and Phenotype 2. Note that this definition is independent of the *p-value* distribution for the phenotypes.

To generate correlated data from the uniform distribution it was necessary to generate correlated data from the normal distribution using Cholesky decomposition and then transform the normal values to uniform using the cumulative distribution function transformation.

The process for simulating GWAS results is described by the following steps:

1) Generated a $j \; x \; 2$ (Z) matrix of two uncorrelated Gaussian random variables $N(0,1)$, where j is the number of SNPs to be compared between the phenotypes.
2) Define the covariance matrix $\sum$, from the required correlation between phenotypes.
3) Find the root of $\sum$, i.e. a matrix C such that $CC^T = \sum$.
4) The target matrix of correlated values is then defined by CZ.
5) Normal distribution values can then be transformed to uniform through the cumulative distribution function.

6.3.2.1 Null model

Under the null hypothesis, either there are no true associations in either GWAS dataset, or the polygenic components are similar between the compared phenotypes. We considered two 'null' scenarios:

1) Neither phenotype GWAS set has any signal of association: *p*-values of association were generated from $N(0,1)$ distributions for both phenotypes.
2) Both phenotypes show a similar signal of association. To approximate this, SNP association *p*-values were generated from ($N(0.1,1)$ distributions for both phenotypes.

6.3.2.2 Alternate model

The alternate hypothesis is that SNPs associated with one phenotype are not associated with the second phenotype. We considered an alternate hypothesis scenario where *p*-values of association were generated from a $N(i,1)$ distribution in the first phenotype and from a $N(0,1)$ distribution for the second phenotype, with *i*=0.1, 0.2 for different strengths of association.

For each simulation model, *p*-value distributions were plotted for each phenotype. Inflation in test statistics were assessed by calculating the genomic control inflation factor, λ, defined as the mean

test statistic on a chi-square distribution. This is a standard measure in genetic studies, and under no inflation, λ=1.

There are many other possible alternative hypotheses that could be modelled here, and different assumptions for the underlying distribution of test statistics within a GWAS study. Here we have focussed on the simplest approach to give an initial indication of the utility of the moth approach for correlated phenotypes. Moth plots were assessed for each simulation, by correlation between phenotype, and characteristic changes in the plot were noted. All the simulation and analyses were conducted using SAS v.9.4 software. Figures were plotted using R software v.3.0.

6.3.3 Application to cancer data

These methods were then applied to genome-wide analysis of the lung and ovarian cancer data from Chapter 5. Associations between the safety phenotypes and SNP allele count were tested using logistic regression. Associations between the efficacy phenotypes and allele dosage were tested using Cox proportional hazard models. Both Cox and logistic models incorporated 5 ancestry principal components, age, gender and cisplatin subclass as covariates (see Chapter 5).

6.4 Results

6.4.1 Simulation study

Under the null hypothesis, histograms for simulated *p*-values with underlying test statistic distributions of $N(0,1)$ and $N(0.1,1)$ in both phenotypes are shown in Figure 6.4.1 and Table 6.4.1, for correlations of 0, 0.5, and 0.9. The histograms confirm that the distribution of *p*-values for both phenotypes is similar and approximately uniformly distributed. Simulating from the $N(0.1,1)$ distribution results in lower *p*-values. This is confirmed by the λ (genomic inflation factor), which is close to one for the $N(0,1)$ simulation but inflated for the $N(0.1,1)$ simulation with λ ≈ 1.1. In each case, the *p*-value correlations across phenotypes are as expected.

Figure 6.4.2 and Table 6.4.2 show the distribution of the simulated *p*-values across varying correlation for both 'alternate' models. The distribution of *p*-values is dissimilar for each phenotype. In both simulations, phenotype 1 has a higher frequency of lower *p*-values while ($\lambda =$

$1.091 \,\&\, 1.190$ for $N(0.1,1) and\ N(0.2,1)$ respectively). By contrast, phenotype 2 exhibits a uniform distribution of *p*-values ($\lambda \approx 1.00$).

The moth plots corresponding to each data simulation, allow us to interpret the impact of test statistic distribution and correlation (Figure 6.4.3). The top-left panel replicates Figure 6.3.2 as it represents a moth plot comparing phenotypes when there is no correlation in the *p*-values between phenotypes and no association (*p*-values for both phenotypes are independent, and from a uniform distribution U(0,1)). Using this panel as a reference we observe the following 3 effects:

1) Increasing correlation between phenotypes reduces the maximum amplitude of *B* & *C* and increases the intercept of the *A*-*D* intersection.
2) Differences in the *p*-value distribution between phenotypes create a difference in the maximum amplitude of *B* and *C* (as phenotype 1 had lower *p*-values, this increases the maximum amplitude of the *B*-function vertex and reduces the amplitude of the ***C***-function vertex, i.e. individual SNP *p*-values are more likely to be below the α-threshold in phenotype 1 and above the α-threshold in phenotype 2).
3) The phenotype pooled mean of *p*-values controls the α value at which *A* and *D* intersect and the vertex of the ***B*** and ***C*** functions.

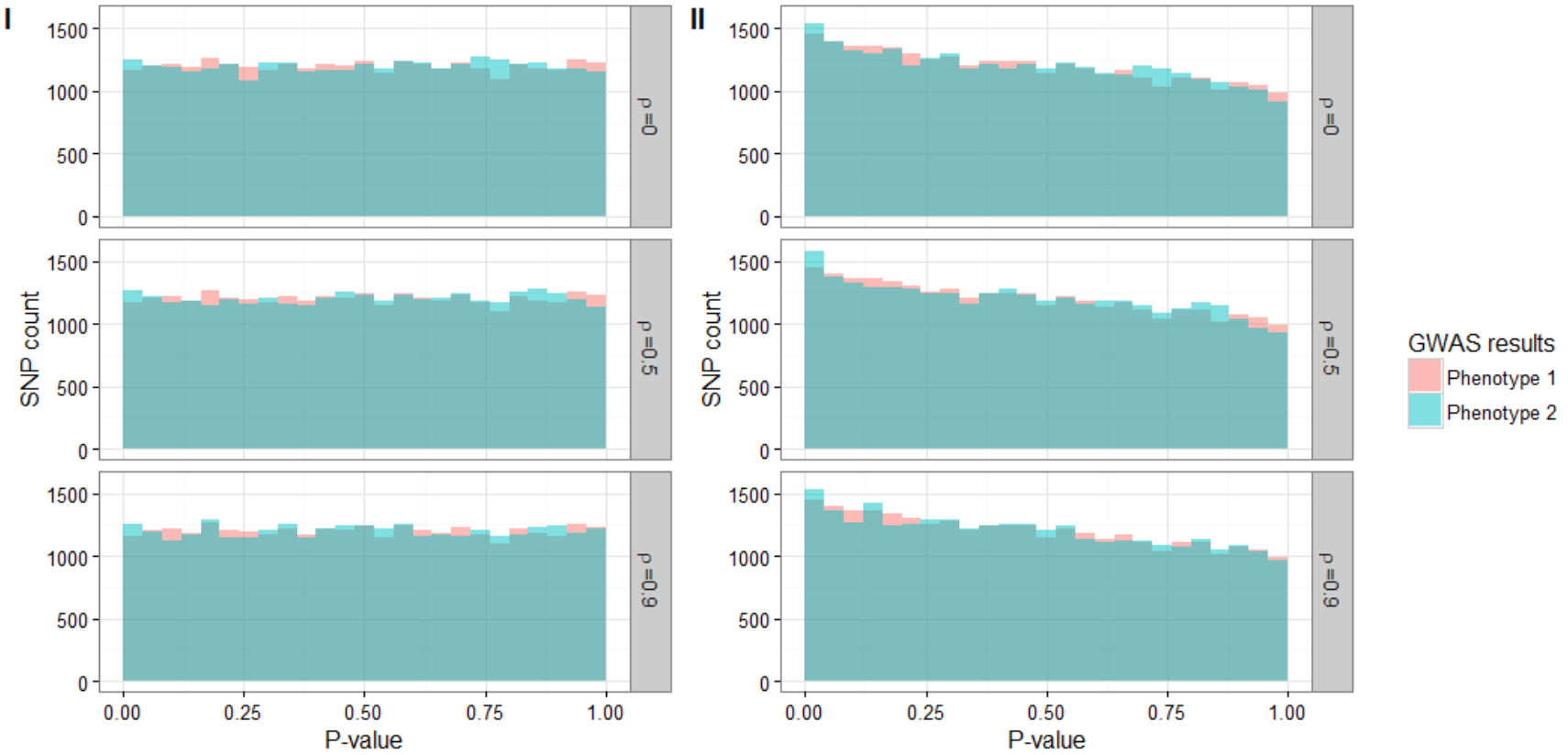


Figure 6.4.1 Null simulations: Histograms showing simulated *p*-value distributions.
In panel **I** the simulated *p*-values for both phenotypes are drawn from $N(0.1,1)$ distributions prior to conversion to $U(0,1)$ through the cumulative distribution function. In panel **II** the simulated *p*-values for both phenotypes are drawn from $N(0.1,1)$. As both phenotypes *p*-values are from the same distributions the frequency of SNPs within each histogram bin is similar both between phenotypes. As expected, panel II shows a higher frequency of SNPs with lower *p*-values.

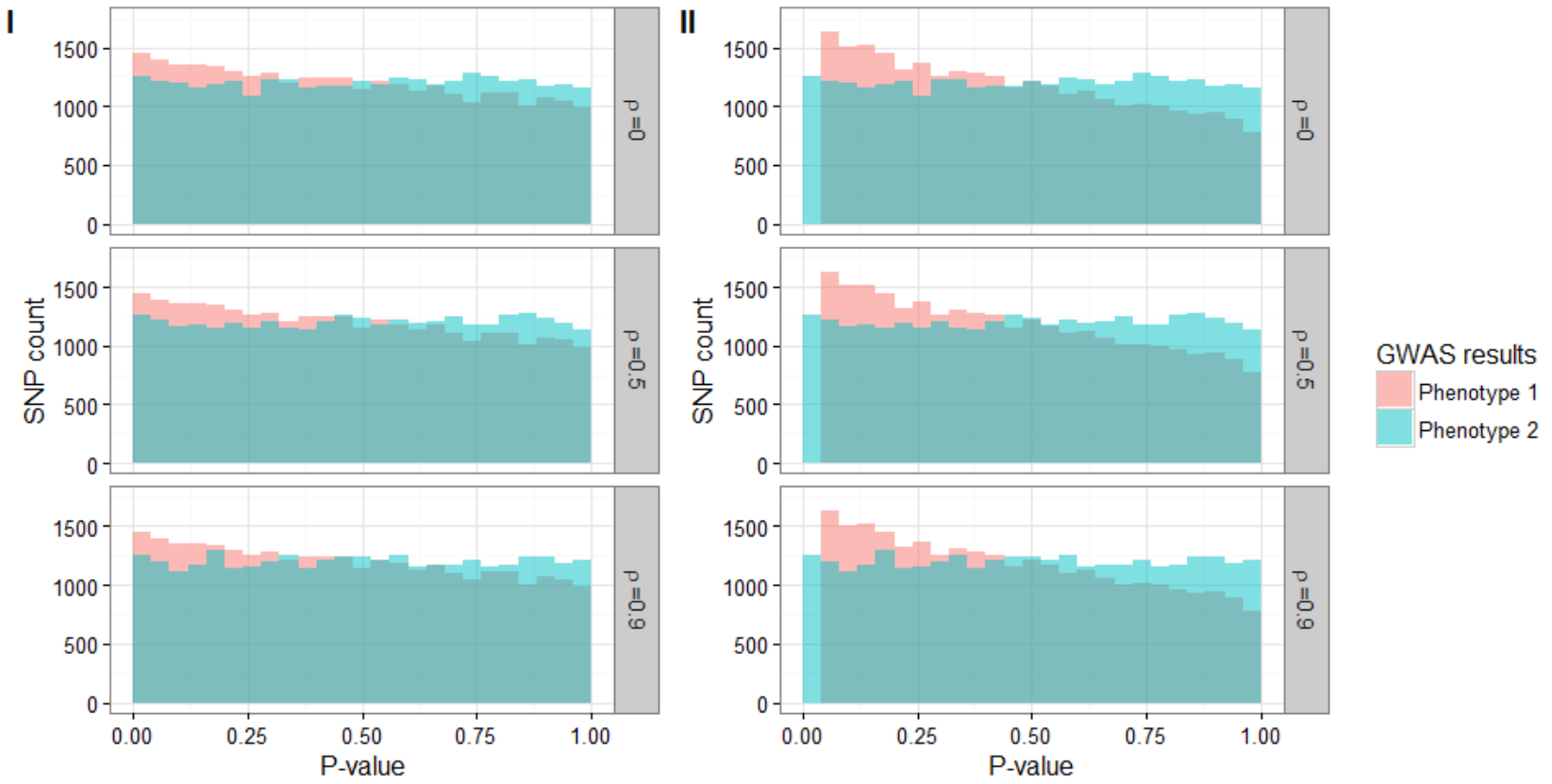


Figure 6.4.2 Alternate simulations: Histograms showing simulated *p*-value distributions.
In both panel I and panel II the *p*-values for Phenotype 2 are generated from a $N(0,1)$ distribution. In panel I the simulated *p*-values for Phenotypes are drawn from $N(0.1,1)$ distribution. In panel II the simulated *p*-values for Phenotype 2 are drawn from $N(0.2,1)$. In both panels Phenotype 1 is biased towards lower *p*-values as compared with Phenotype 2.

Table 6.4.1 Descriptive statistics for simulated *p*-values under null hypothesis models

Sample Distribution	Correlation	Variable	Mean (μ)	Lambda (λ)	Median	Minimum	Maximum	Spearman Correlation Coefficient
Phenotype 1 & 2 $N(0,1)$	0	Phenotype1	0.49952	0.998	0.49695	0.0000724	0.99996	0.00024
		Phenotype2	0.50102	1.006	0.50413	0.0000844	0.99997	
	0.5	Phenotype1	0.49952	0.998	0.49695	0.0000724	0.99996	0.48192
		Phenotype2	0.50138	1.010	0.50294	0.0000165	0.99993	
	0.9	Phenotype1	0.49952	0.998	0.49695	0.0000724	0.99996	0.89108
		Phenotype2	0.50026	1.003	0.49987	0.0000207	0.99998	
Phenotype 1 & 2 $N(0.1,1)$	0	Phenotype1	0.47134	1.091	0.45714	0.0000482	0.99993	0.00024
		Phenotype2	0.47283	1.100	0.46429	0.0000563	0.99995	
	0.5	Phenotype1	0.47134	1.091	0.45714	0.0000482	0.99993	0.48192
		Phenotype2	0.47319	1.103	0.46310	0.0000106	0.99990	
	0.9	Phenotype1	0.47134	1.091	0.45714	0.0000482	0.99993	0.89108
		Phenotype2	0.47207	1.097	0.46004	0.0000133	0.99996	

Table 6.4.2 Descriptive statistics for simulated *p*-values under alternate hypothesis models

Sample Distribution	Correlation	Variable	Mean (μ)	Lambda (λ)	Minimum	Maximum	Spearman Correlation Coefficient
Phenotype 1: $N(0.1,1)$ Phenotype 2: $N(0,1)$	0	Phenotype1	0.47134	1.091	0.0000482	0.99993	0.00024
		Phenotype2	0.50102	1.007	0.0000844	0.99997	
	0.5	Phenotype1	0.47134	1.091	0.0000482	0.99993	0.48192
		Phenotype2	0.50138	1.010	0.0000165	0.99993	
	0.9	Phenotype1	0.47134	1.010	0.0000482	0.99993	0.89108
		Phenotype2	0.50026	1.003	0.0000207	0.99998	
2henotype 1: $N(0.2,1)$ Phenotype 2: $N(0,1)$	0	Phenotype1	0.44331	1.190	0.0000317	0.99990	0.00024
		Phenotype2	0.50102	1.010	0.0000844	0.99997	
	0.5	Phenotype1	0.44331	1.190	0.0000317	0.99990	0.48192
		Phenotype2	0.50138	1.010	0.0000165	0.99993	
	0.9	Phenotype1	0.44331	1.190	0.0000317	0.99990	0.89108
		Phenotype2	0.50026	1.003	0.0000207	0.99998	

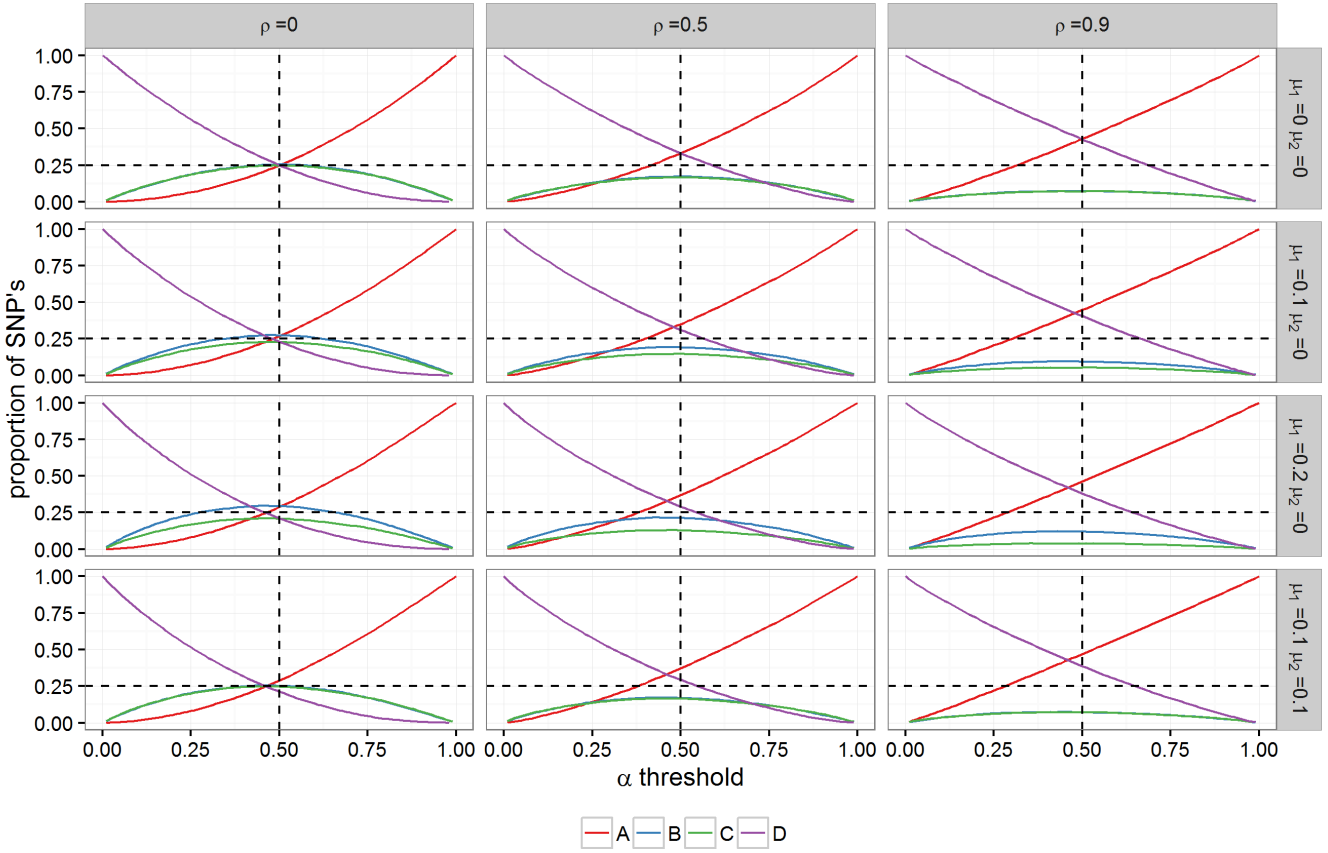


Figure 6.4.3 Moth plots for all of the simulation scenarios.
The first and last rows plot the null simulations (Phenotype 1 ~ $N(0,1)$ = Phenotype 2 ~ $N(0,1)$ and Phenotype 1 ~ $N(0.1,1)$ = Phenotype 2 ~ $N(0.1,1)$ respectively). The second and third rows contain the alternate simulations (Phenotype 1 ~ $N(0.1,1) \neq$ Phenotype 2 ~ $N(0,1)$ and Phenotype 1 ~ $N(0.2,1) \neq$ Phenotype 2 ~ $N(0.1,1)$ respectively).

6.4.2 Cancer data – efficacy and safety phenotypes

The genetic summary statistics from the ovarian and lung cancer cohorts were explored using the moth plots for the four primary phenotypes previously assessed: survival endpoints of OS and PFS, and safely endpoints of GID and NEU.

6.4.2.1 Interpreting the data from the moth plots

The moth plots for each pair-wise combination of these endpoints, split by cohort, are shown in Figure 6.4.4. Using the information from section 6.4.1 we assert the following from the figure:

Lung cohort

1) In the comparison of PFS and OS
    a. PFS and OS are correlated as indicated by **B** and **C** vertices being lower than the **A-D** intercept.
    b. The **B** and **C** overlap, indicating that the distribution of *p*-values is similar between phenotypes
    c. The **A-D** intercept is below 0.5 indicating that the distributions of OS and PFS are biased towards lower *p*-values.
2) In the comparison of OS vs. safety endpoints – (GID and Neu) there is slight separation between **B** and **C** (with **B** slightly higher) indicating that the *p*-values are higher in the safety endpoints and that OS does not have the same distribution of *p*-values as the safety endpoints.
3) With the exception of OS and PFS, the vertex of **B-C** meets the height of the intercept of the **A-D** functions implying that apart from OS and PFS, the correlation between phenotypes is close to zero.

Ovarian cohort

4) In the comparison of PFS and OS
    a. **B** and **C** vertices are lower than the **A-D** intercept indicating that PFS and OS are correlated.
    b. **B** and **C** overlap, this indicates that the distribution of *p*-values is similar between phenotypes
    c. The **A-D** intercept is below 0.5 indicating that the distributions of OS and PFS are biased towards lower *p*-values.
5) In the comparison of OS vs. safety endpoints – (GID and Neu) there is slight separation between the **B** and **C** functions (**B** slightly higher), indicating that the *p*-values are not uniformly distributed in the safety endpoints, and also that OS does not have the same distribution of *p*-values as the safety endpoints.
6) In the comparison of PFS vs. safety endpoints – (GID and Neu) there is separation between the **B** and **C** (**B** slightly higher) indicating that the *p*-values are higher in the safety endpoints and also that PFS does not have the same distribution of *p*-values as the safety endpoints.
7) With the exception of OS and PFS, the four curves, A, B, C, D intersect at (0.5,0.25) as seen under the null hypothesis in the simulated data, implying that the correlation between these phenotypes is close to zero.
8) In the comparison of the safety endpoints (GID and Neu) the *A-D* intercept is above α=0.5 on the x-axis implying that the mean of *p*-values across both phenotypes is greater than 0.5.

Between cohorts

9) The vertices of **B-C** are lower in the lung cohort as compared with the ovarian cohort indicating the correlation between OS and PFS is stronger in the lung cohort.
10) The separation between **B-C** vertices in the comparison of efficacy and safety endpoints is greater in the ovarian cohort as compared with the lung cohort implying that the differences in distribution between efficacy and safety endpoints is more pronounced in the ovarian cohort data.

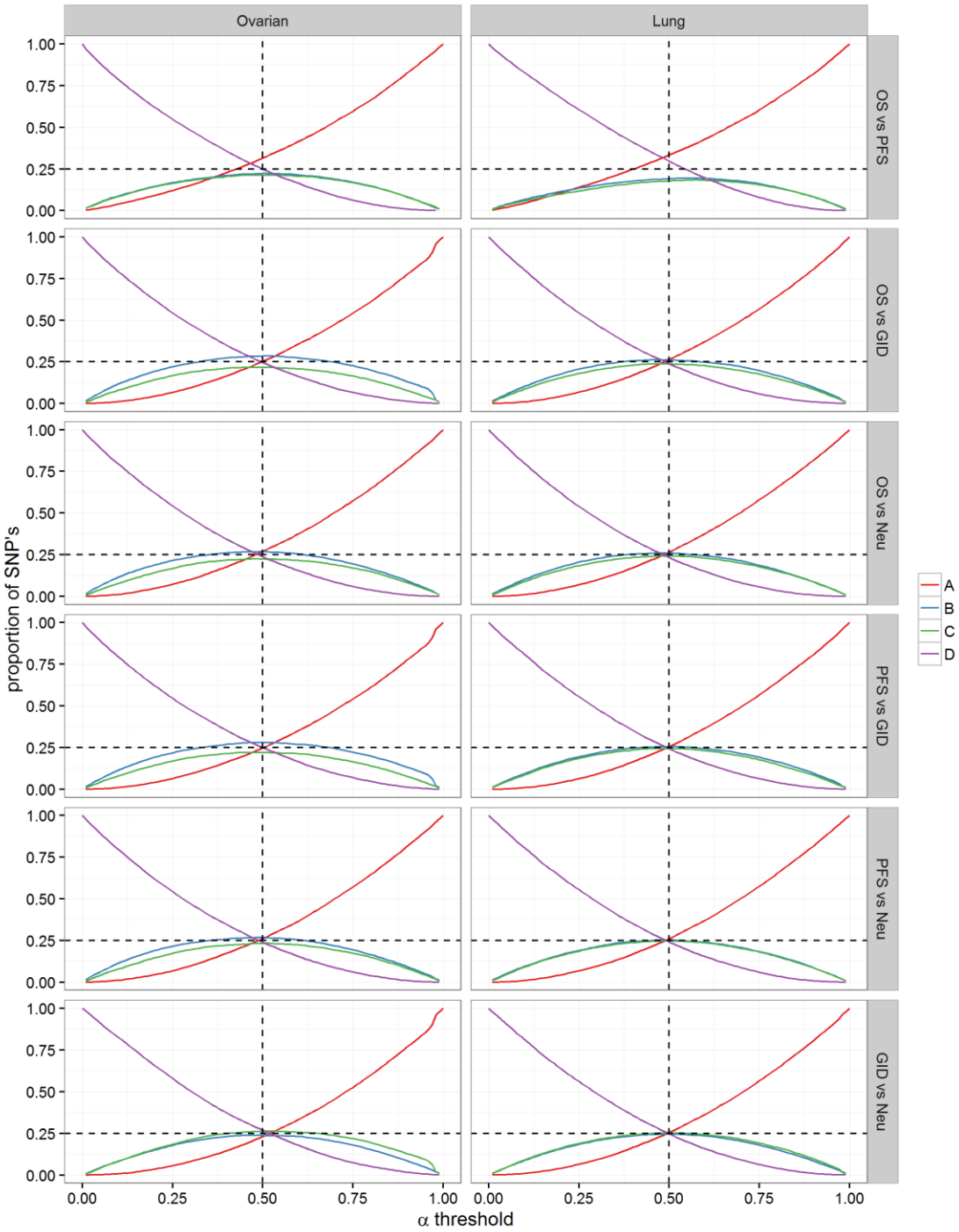


Figure 6.4.4 Moth plots comparing each phenotype by cohort.
OS=overall survival, PFS=progression free survival, GID=gastrointestinal disorder, Neu=Neutropenia.

We can assess the validity of our assertions using a series of exploratory data analysis techniques for each phenotype, including summary statistics, histogram plots and correlations. Figure 6.4.5 shows the distribution of *p*-values for all four (two efficacy: PFS and OS; and two safety: GID and Neu) cancer endpoints by cohort; Table 6.4.3 presents the descriptive statistics for each endpoint and Table 6.4.4 presents the correlation between endpoints.

Figure 6.4.5 shows that our assertions regarding the distributional differences between efficacy and safety endpoints and between cohorts are upheld. Both efficacy endpoints in both cohorts show systematic bias towards lower *p*-values (μ = 0.474, 0.478, 0.482 and 0.491 for OS and PFS endpoints in the ovarian and lung cohorts respectively) (similar in distribution to Phenotype 1 from our alternate simulations). Moreover, this shift is more pronounced in the ovarian cohort as compared with the lung cohort (λ = 1.16, 1.14, 1.09 and 1.04 for OS and PFS endpoints in the ovarian and lung cohorts respectively). Both safety endpoints in both cohorts are approximately uniformly distributed, similar in shape to the null simulations with $N(0,1)$)(μ = 0.525, 0.506, 0.502 and 0.495; λ = 0.952, 0.948, 0.994 and 0.994) for GID and Neu endpoints in the ovarian and lung cohorts respectively).

Table 6.4.4 confirms that OS and PFS are the only two endpoints exhibiting a correlation markedly different from zero and that the correlation is greater in the lung cohort (ρ=0.2014 and ρ=0.3872 for the ovarian and lung cohorts respectively).

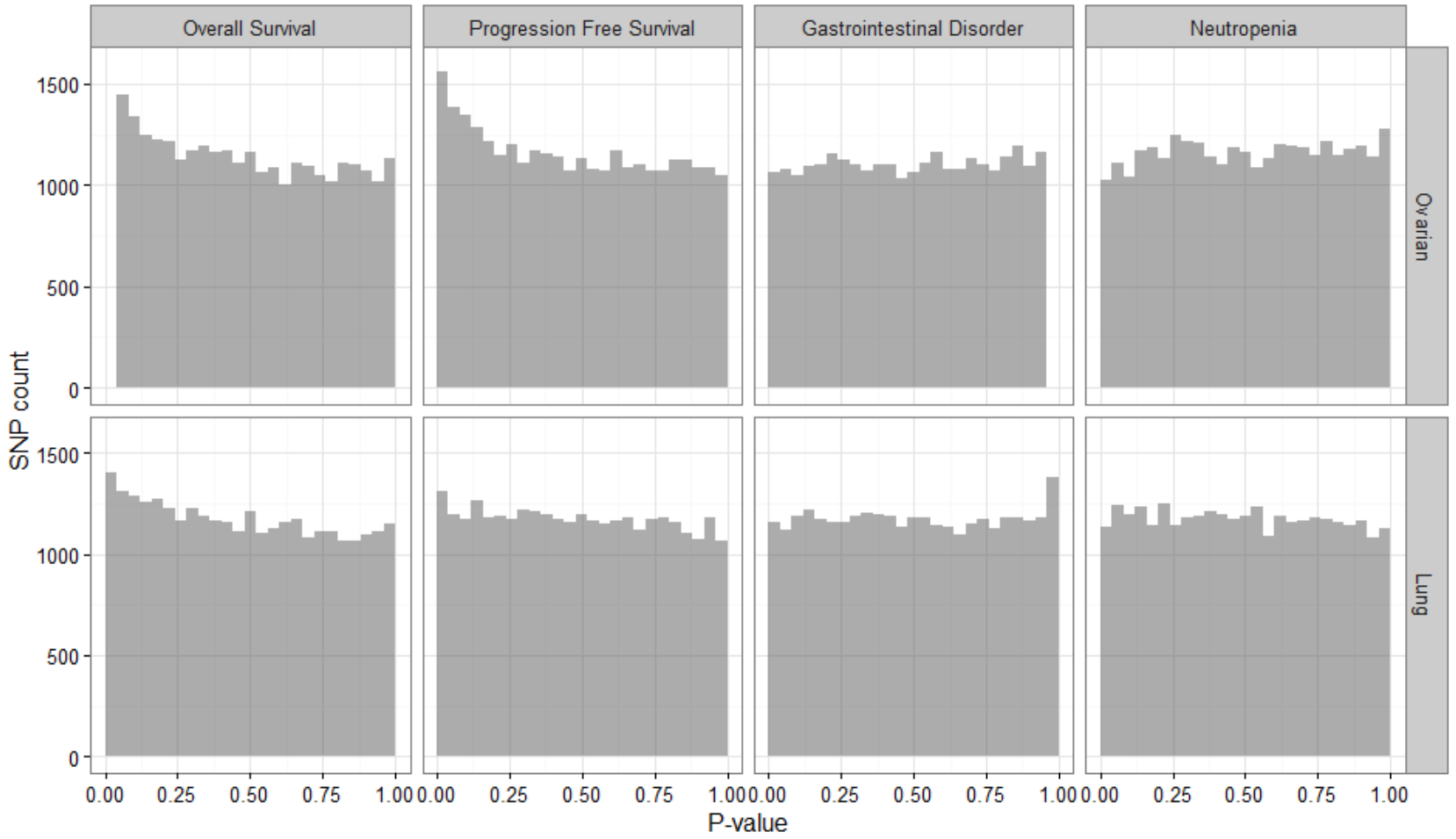


Figure 6.4.5 The histograms show the distribution of *p*-values for the association tests of each phenotype by cohort.

Table 6.4.3 Descriptive statistics for the efficacy and safety endpoints.

Cohort	N	Variable	Mean (μ)	Minimum	Maximum	Lambda (λ)
Ovarian	28961	OS	0.474	1.13E-06	0.999914	1.16026
		PFS	0.478	3.54E-06	0.999991	1.14258
		GID	0.527	1.56E-05	0.999945	0.95176
		NEU	0.506	3.15E-05	0.999983	0.94792
Lung	29258	OS	0.482	2.23E-06	1	1.09337
		PFS	0.491	3.52E-06	0.999917	1.04988
		GID	0.502	4.09E-05	0.999998	0.99354
		NEU	0.495	3.57E-05	0.999956	0.99383

Table 6.4.4 Correlation between *p*-values from alternate phenotypes

Spearman Rank Correlation				
	OS	PFS	GID	NEU
		Ovarian		
OS	1	0.204	0.003	0.022
PFS	0.387	1	-0.001	0.012
GID	0.000	0.004	1	0.001
NEU	-0.001	0.009	0.010	1
	Lung			

6.5 Discussion

We introduce a framework exploring the similarities and differences of genetic association results from individual phenotype analyses. We have shown that it is possible to capture in a single graphic both correlation and distributional differences between results from two phenotypes, and this moth plot effectively combines the information from histograms, summary statistics and correlation tables.

The value of this method is to provide an easily produced graphic that can summarise several aspects of the concordance/discordance between univariate genotype-phenotype association results. In GWAS testing the association between genetic variants and phenotypes, each one-at-time, has remained the method of choice, even in studies where multiple phenotypes are collected. Such studies typically report the subset of SNPs achieving genome wide significance by phenotype and discuss the similarity or difference in this subset [270] without exploring the wider implications of the concordance/discordance (e.g. [271]). More recently, methods have been developed to conduct association analysis of multiple traits in a genome-wide association study – a multivariate GWAS [266, 272]. The multivariate methods provide a measure of the shared genetic variance through a correlation analysis between genetic portions of variability (where significant genetic correlation indicates shared genetic variance). While these methods have been demonstrated to provide greater power to detect genotype-phenotype associations (and pleiotropic effects) again researchers are left discussing the significant results relative to the univariate analyses. Consequently, our method of comparing GWAS results is applicable to both historical and modern methods offering a means of interrogating the genome-wide data when joint analyses methods do not yield replication or pleiotropy.

Currently, our descriptions of the moth plots are qualitative. This method could be expanded to provide quantitative descriptions matching λ and ρ parameters through employing non-linear regression techniques to provide accurate estimates of the ***B*** and ***C*** vertices along with the co-ordinates of the A-D intercept. Quantification of key function parameters could potentially allow for the formulation of single metric parameters that would further simplify the interpretation of the moth plots, or the derivative metric could even replace the moth plot. More importantly, a nonlinear regression may offer a way of mapping ‘significant’ differences between ***B*** and ***C***

vertices to ‘significant’ differences in the mean of the study results being compared. Currently, the qualitative nature of this method does not offer this link.

The widespread use of ‘statistical significance’ as a definite measure of a scientific finding has degraded *p*-values into ‘significant’ and ‘non-significant’ categories [273]. Within this binary interpretation, statistical tests are generally called significant, with rejection of the null hypothesis, if the *p*-value falls below a predefined alpha value. It is therefore implicit that any significant result greater than alpha is non-significant no matter how close or far away to the critical threshold.

This binary interpretation of significance results protects against an increased prevalence of scientific findings that provide at best, weak evidence against the null hypothesis and consequently represent the consensus viewpoint of medical researchers and biostatisticians [274].. While we support this perspective in the identification to disease associated variants, our method treats each signal (irrespective of its distance from genome wide significance) as equally informative and continues the idea of using *p*-value as a continuous measure of evidence against the null rather than as a binary decision rule [275-279].

Genome-wide association studies classically focus on the small subset of SNPs that achieve significance at the stringent significance level of 5e-08. This threshold reflects traditional nominal significance with this *p*-value achieved by chance in only 5% of null genome-wide studies. However, many SNPs with *p*-values below this threshold may still be associated with the phenotype, when the trait is highly polygenic and the study has limited power. Other analysis methods use a similar genome-wide strategy to the moth plots, supporting our hypothesis that genome-wide information, regardless of *p*-value, gives valuable insights. For example, polygenic risk scores (PRS) combine risks across SNPs to create a measure of genetic liability for each individual. These risk scores can be constructed including only genome-wide SNPs, but extensive experience across disorders shows that including SNPs with more liberal *p*-values increases the predictive ability of these scores. Indeed, many studies include all SNPs in constructing PRS, regardless of association *p*-value.

Gastrointestinal disorder (nausea, vomiting) and neutropenia are common adverse events associated with the use of platinum therapy in both lung cancer [280, 281] and ovarian cancer [98, 119] patients. In our data, a substantial proportion of patients experienced both

gastrointestinal and neutropenic events across both cohorts during their treatment. As these two AEs often occur in conjunction, we hypothesised that the phenotypic correlation would arise from an underlying genetic correlation. We expected to find a moderate correlation in the genetic association signals between these two safety phenotypes. It is therefore surprising that the association signals exhibited such low correlation across both cohorts of data. With the caveat that there is only limited power to detect SNP associations in these cohorts, we might cautiously interpret these results as evidence of low genetic correlation between the variants that give rise to each phenotype. Alternatively, as both gastrointestinal disorders and neutropenia have symptoms that can be managed through the use of concomitant medications it is possible that shared genetic architecture is being masked as not all susceptible patients are exhibiting the phenotype. Of course, the simplest explanation that none of the SNPs tagged in our study are associated with either safety phenotype. This would explain both the lack of significant results and the low genetic correlation. The exome-wide genotype data for these cohorts will also have limited our power to detect genetic correlation, which might be detectable with full genome-wide genotyping.

The low correlation between efficacy and safety association results fails to support our earlier hypothesis (Chapter 3) that safety and efficacy are linked. If adverse event sensitivity is conferred through genetic variation, then variant-carrying patients should be more likely to experience DLTs and consequently have poorer survival outcomes. The strongest observed genetic correlation was between overall survival and progression free survival, but the evidence was still low despite the high phenotypic correlation between the endpoints of overall survival and progression free survival. Within lung cancer, very few factors ameliorate the post-progression decline in patient health and therefore the post-progression time to death should not vary greatly between patients. This means that time to progression and time to death should have similar distributions and therefore high correlation. Moreover, a subset of patients will experience undetected progression events and their PFS time will be identical to their OS time ($r_s = 0.800$; Spearman rank Correlation (r_s)of patient patients experiencing both OS and PFS events). The low correlation between genetic association signals between these two intertwined phenotypes highlights the challenges in interpreting genetic results. The definition of separate phenotypes is largely based upon historical clinical perspectives and may not be grounded in current biological perspectives. A strong correlation between variants associated with OS and variants associated with PFS would

need to be interpreted as evidence of variant pleiotropy arising from similarity of phenotypes. The low correlation observed supports the perspective that only association *p*-values below a reasonable alpha threshold can be used to assess genetic correlation.

Despite these limitations, the moth plot may be a useful exploratory data analysis addition to the epidemiological toolbox, since it allows rapid, easily interpretable interrogation of concordance/discordance of association results.

Chapter 7. Sample size calculations for repeated measures experiments – a review

7.1 Abstract

Estimating the required sample size is an integral step in study design [282]. Methods to calculate the required sample size for cross sectional experimental designs are well understood and supported by widely available software [283]. However, advances in analysis methods for repeated measures data have left a gap between the practice of the sample size calculation during the planning of clinical trials and the complexity of the experimental designs and analysis methods most frequently employed. In this chapter, we compile available sample size calculations for continuous outcomes of repeated measures data, with a baseline score and at least one follow-up measure collected on each participant. We include three extensions to calculate the required sample size when: (1) the baseline score is part of the outcome vector, (2) baseline score is considered as a covariate; and (3) the analysis is based upon change from baseline scores.

7.2 Introduction

Much of medical research aims to estimate the causal relationships between interventions and patient outcomes. Randomised Controlled Trials (RCTs) have become one of the main tools that researchers use to achieve this aim [284-286].

Cross sectional RCTs are usually set up with the objective of estimating the impact of a novel intervention in a sample of subjects at a single time point. By contrast, repeated measures RCTs track the trajectory of response (change-over-time or slope of response) to tested interventions by collecting multiple measurements from each subject across time. Repeated measurement designs are generally considered more efficient for determining a treatment effect as compared with cross sectional designs; collecting repeated measurements can increase statistical power for a fixed number of subjects [287].

When setting up any RCT, one of the most critical tasks is to calculate the sample size that will be used for the trial [288]. This is to ensure that the planned sample is large enough to detect

clinically relevant differences in outcomes between comparator groups (treatment and control group). A sample size that is too small leads to an underpowered study, which will have a high probability of failing to reject the null hypothesis, even when the intervention has a clinically meaningful effect. Overpowered studies, with samples larger than required, may potentially expose a larger pool of individuals to an untested treatment; these studies will be more logistically complex, potentially take longer to complete and delay conveying important information to clinicians and patients [289]. Consequently, the optimal sample size for a trial is a 'goldilocks' problem that requires balancing the number of patients to ensure a sufficient number to provide adequate power while minimising the number of patients exposed to an inferior treatment [289, 290].

Despite the gain in efficiency associated with repeated measures designs, there are several complications in determining the appropriate sample size. Repeated measurements taken from the same subject tend to be correlated, and the correlations must be accounted for in calculating the sample size. Failure to account for intra-patient correlation between the repeated measurement outcomes, or incorrect specification of the correlation, can result in erroneous sample size estimates. For the analysis of repeated measures data, both generalised linear mixed-effects models and marginal models provide robust analysis tool to deal with subject correlation of response across repeated measures.

There are two aspects of the intra-correlation that form part of a robust repeated measures sample size calculation: 1) the correlation strength, and 2) the correlation structure. The correlation strength and structure among repeated measurements within a subject can be estimated from previous studies, a pilot study, or an educated speculation based on investigator's experience.

A further complication to repeated measures sample size calculation is introduced by the parameterization of the baseline measurement. As inclusion of the baseline as a predictor or response has the potential to alter the statistical efficiency of the analysis method, it should be accounted for in the sample size calculation.

7.2.1 Some general notation

To aid in the statistical discussion below it is helpful to present the notation for various aspects of longitudinal design. We index the N subjects in the longitudinal study as:

$$i = 1, \ldots, N$$

For *G+1* treatment groups, the total number of patients can be partitioned into the subset of patients with each treatment group (n_j) such that:

$$\sum_{j=0}^{G} n_j = N$$

For a balanced study design in which all subjects have complete data, and are measured at the same time points, such that all subjects have an equal number of observations, we index the assessment occasions as:

$$k = 0, \ldots, T \text{ observation time points}$$

Such that the total number of time points is equal to $T + 1 = m$. Additionally, we use k and l to denote two separate time points.

The repeated responses, or outcomes, or dependent measures for subject *i* are denoted as the vector:

$$Y_{ijk} = \begin{pmatrix} Y_{ij0} \\ \vdots \\ Y_{ijT} \end{pmatrix} \text{ all measurements for subject } i$$

Such that:

$$\bar{Y}_{\ldots} = \bar{Y} = \frac{\sum_{k=0}^{T} \sum_{i=1}^{N} Y_{ijk}}{Nm} \text{ is the overall mean}$$

$$\bar{Y}_{\cdot jk} = \mu_{jk} = \frac{\sum_{i=1}^{n_j} Y_{ijk}}{n_j} \text{ is the mean for group } j \text{ at time } k$$

7.3 Methods

7.3.1 Pre-post designs

A pre-post design is a repeated measures experiment in which all the subjects are measured at baseline (pre-treatment), and then once again at some time point after (post-treatment). The pre-

treatment data are intended to measure any differences that existed before the treatment administration while the post-treatment data are used to test if the treatment influenced the response variable. In general there are three non-repeated measures methods commonly used to analyse pre-post data [291]:

1) Use only the post-treatment assessment ignoring the pre-treatment responses: **POST**;
2) Use the patient specific change scores (post-treatment minus pre-treatment): **CSA**;
3) Use the baseline/pre-treatment scores as a covariate along with treatment with the post-treatment measurement as the response variable: **ANC**.

Overall and Starbuck [292] and Overall [293] explored how to choose the sample size for pre-post design experiments, with consideration for how the baseline measurement is incorporated into the analysis. The sample size formulae presented in sections 7.3.1.2 and 7.3.1.3 are based upon their work.

Similar to cross sectional studies, we need the following information to obtain sample size estimates for pre-post repeated measurement studies:

1) The type I error (α)
2) The power $(1-\beta)$
3) The clinically meaningful difference between two treatment groups estimated as the difference between group means at the post-baseline measurement $(\mu_{11}-\mu_{01})$. Where μ_{01} and μ_{11} represent the means at the post-treatment measurement for the control (0) and active (1) treatment groups respectively.
4) The variance of the post-baseline responses (σ^2)

Additionally, due to the repeated measures aspect, depending on how the baseline measurement is used, pre-post sample sizes also make use of:

5) The correlation between pre-treatment and post-treatment responses across both treatment groups (ρ)

7.3.1.1 POST

In a randomised clinical trial, pre-treatment means are expected to be approximately equal across treatment arms. Consequently, the baseline measurement is often ignored and the mean difference between treatment groups at the post-treatment measurement can be used to compare the treatment effect [291] using a two-sample mean test.

For this type of analysis, the sample size required for a given level of alpha and power is estimated using methods appropriate for a cross-sectional study [294, 295]:

$$n = \frac{2\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_{11} - \mu_{01})^2} \quad (7.3.1)$$

Where $z_{1-\frac{\alpha}{2}}$ and $z_{1-\beta}$ are the $100\left(1-\frac{\alpha}{2}\right)th$ and $100(1-\beta)th$ percentiles of the standard normal distribution, *n* is the sample size per group. Equation (7.3.1) can be rearranged to solve for power given a fixed sample size:

$$z_{1-\beta} = \sqrt{\frac{n}{2}}\frac{(\mu_{11} - \mu_{01})}{\sigma} - z_{1-\frac{\alpha}{2}} \quad (7.3.2)$$

7.3.1.2 CSA

While it is reasonable to assume that pre-treatment / baseline values will be balanced, chance variation can often result in some doubt concerning the true treatment effect and it is therefore not uncommon that investigators will use the change from baseline as the outcome of primary interest on which to compare treatment groups.

The sample size for testing the difference in change scores between two treatment groups can be obtained using the means at the post-treatment $(\mu_{11} - \mu_{01})$, the common error variance (σ^2) estimated from either the baseline or final assessment time point, and the pre-post assessment correlation (ρ). Owing to the regression-to-the-mean phenomenon [296, 297], the variance of the difference score between the first and the last measurements can be written as [295]:

$$\sigma_{CSA}^2 = 2(1-\rho)\sigma^2 \quad (7.3.3)$$

$$n = \frac{2\sigma_{CSA}^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_{11} - \mu_{01})^2} = \frac{4(1-\rho)\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_{11} - \mu_{01})^2} \quad (7.3.4)$$

Equation (7.3.4) shows that the sample size for testing difference scores between two treatment groups is equal to that for testing the significance between the raw post-treatment scores

multiplied by $2(1-\rho)$. This means that when the correlation between pre- and post-measurements is less than 0.5, the use of change scores will result in an inflated sample size as compared with using the raw post-treatment scores. Conversely, when the correlation between pre- and post- measurement is greater than 0.5, the use of change scores will give greater power for a fixed sample size than using the raw post-treatment scores.

The power of the analysis of change scores for a fixed sample size can be obtained from[295]:

$$z_{1-\beta} = \frac{(\bar{Y}_{11} - \bar{Y}_{01})}{\sigma}\sqrt{\frac{n}{4(1-\rho)}} - z_{1-\frac{\alpha}{2}} \tag{7.3.5}$$

7.3.1.3 ANC

Entering the baseline score as a covariate to adjust for baseline differences is frequently heralded as the gold standard for the analysis of repeated measures data. The inclusion of baseline score as a covariate can also be applied to difference score analysis. Both cases will yield identical results. If ρ is the correlation between pre-treatment and post-treatment assessment time points then the proportion of variance accounted for by the baseline covariate is equal to ρ^2. The error variance remaining after removing the variance associated with the pre-treatment score is therefore [295]:

$$\sigma_{ANC}^2 = (1-\rho^2)\sigma^2 \tag{7.3.6}$$

Where σ^2 is the common within subject variance. The sample size estimate per group for testing either the post-treatment means, or change scores, with baseline entered as a covariate is [295]:

$$n = \frac{2\sigma_{ANC}^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_{11}-\mu_{01})^2} = \frac{2(1-\rho^2)\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_{11}-\mu_{01})^2} \tag{7.3.7}$$

Equation (7.3.7) shows that the sample size for testing either the post-treatment means or difference scores between two treatment groups, after adjusting for baseline, is equal to that for testing the significance between the raw post-treatment scores multiplied by $(1-\rho^2)$. The adjustment that is made is the correlation between the covariate and the response variable [288]. Therefore, for any value of $\rho > 0$, adjusting for baseline will result in greater power for a fixed

sample size as compared with analysing only the raw post-treatment measurements. Similarly, inclusion of the baseline as a covariate will result in greater power for a fixed sample size as compared with the analysing the change scores for any value of $\rho < 1$.

Overall and Starbuck provide a detailed discussion on the differences between including or excluding baseline as a covariate in the analysis of change scores. Within this discussion they note that the ratio of efficiency between exclusion of baseline vs inclusion of baseline is equal to [295]:

$$\frac{2(1-\rho^2)}{4(1-\rho)} = \frac{2(1-\rho)(1+\rho)}{4(1-\rho)} = \frac{(1+\rho)}{2} \tag{7.3.8}$$

Equation (7.3.8) can be interpreted as: the sample size for testing difference scores between two treatment groups without baseline score adjustment is equal to that for testing the significance of the difference between treatment groups with baseline score adjustment multiplied by $2/(1+\rho)$. This means that for all values of $\rho < 1$, for a fixed level of power, the required sample size when using baseline as covariate will be less than the sample size required if change scores are used to analyse the data. Interestingly the formula also indicates that when the correlation between the pre- and post-measurements is zero, using baseline as covariate offers almost twice the efficiency of analysing the data using change scores.

Rearrangement of equation (7.3.7) yields the power calculation for a fixed sample size using the baseline as a covariate:

$$z_{1-\beta} = \sqrt{\frac{n}{2}}\frac{(\mu_{11}-\mu_{01})}{\sigma\sqrt{1-\rho^2}} - z_{1-\frac{\alpha}{2}} \tag{7.3.9}$$

7.3.2 Multi-follow-up repeated measures (beyond pre-post)

We use the term multi-follow-up repeated measures to describe a repeated measures experiment in which patients are assessed at baseline, randomised, and then undergo more than one assessment during the treatment period. Overall and Doyle (1994) provided sample size formulas

for repeated measurement designs and Overall (1996) used simulation to examine the effects of different sample sizes. The equations presented below are adapted from their work.

Sample size calculations of multi-follow-up experiments require inputs that differentiate them from the sample size methods for pre-post studies, namely:

(1) the contrast vector appropriate for the hypothesis being examined (a) (see section 7.3.2.1)
(2) the mean difference between treatment groups can no longer be represented by a single value and instead the difference between treatment groups at each time point, is quantified with a vector (d)
(3) the intra-subject correlation matrix for the repeated measurements (C)

7.3.2.1 Orthogonal polynomial contrasts

Researchers examining repeated measures RCT data might be interested in whether the response pattern across time is linear, quadratic, cubic, and so on. The treatment difference across time will then be analysed based upon the observed pattern of response, i.e. whether treatments differ with regard to linear, quadratic or cubic trend. This treatment comparison is achieved using orthogonal polynomial components for time. Within the context of repeated measures experiments, the phrase orthogonal polynomial refers to a coding of time such that polynomial components are statistically independent of one another. In the special case of a linear trend over time, the slope of the regression line appropriately summarises the rate of change of outcomes over time. Again assuming a linear trend over time, the difference in longitudinal rates of change (slopes) between treatment groups is captured by the interaction between the treatment and time from the model estimates [298]. For each subject, rates of change or trends can be constructed by applying appropriately chosen weighting coefficients to the repeated measurements. Table 7.3.1 presents the orthogonal contrast coefficients across a range of *m* used to explore whether the rates of changes across time are the same among treatment groups or whether there is no significant interaction effect between time and treatment group.

Table 7.3.1 Orthogonal coefficients for a linear trend

Total Number of assessment time-points ($m = T + 1$)	**Contrast Coefficients (a_k)**						$\sum_{k=0}^{T} a_k = 0$	$\sum_{k=0}^{T} a_k^2$
	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$		
2	-1	+1					0	2
3	-1	0	+1				0	2
4	-3	-1	+1	3			0	20
5	-2	-1	0	+1	+2		0	10
6	-5	-3	-1	+1	+3	+5	0	70

Sample size estimates for testing a difference in the rates of change can be obtained by using linear trend scores that are constructed by applying linear orthogonal polynomial coefficients to the repeated measurements. The linear orthogonal polynomial coefficients are also called linear contrast coefficients. The linear orthogonal polynomial coefficients (a_j) are equally spaced, and sum to zero:

$$\sum_{k=0}^{T} a_k = 0 \tag{7.3.10}$$

The difference between linear trend means for two treatment groups can be calculated by applying the linearly weighted coefficients to the differences between treatment means at each measurement time point:

$$\sum_{k=0}^{T} a_k d_k = \sum_{k=0}^{T} a_k(\mu_{1k} - \mu_{0k}), \tag{7.3.11}$$

Where d_k is the difference between treatment means at the k^{th} $(k = 0, \dots, T)$ measurement time, which is equivalent to the difference in the observed means for treatment groups 1 (μ_{1k}) and 0 (μ_{0k}) at the k^{th} measurement time.

7.3.2.2 Correlation structures

The correlation matrix (C) describes the variance adjusted covariance matrix for patient errors (e) across the assessment time points. That is the vector e_{ijk}, defined as $e'_{ijk} = (e_{ij0,}e_{ij1,\ldots,}e_{ijT})$ is assumed to be multivariate normal, $MVN(0, R)$, where R is the covariance matrix amongst the e_{ijk} within the i^{th} subject.

Three common covariance models for correlated errors are:

Independent errors $$R = \mathrm{I}\sigma^2$$

Compound Symmetry (CS) $$R_i = \sigma^2 C = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Autoregressive (AR) $$R_i = \sigma^2 C = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & 1 \end{bmatrix}$$

Unstructured (UN) $$R_i = \sigma^2 C = \sigma^2 \begin{bmatrix} 1 & \rho_{01} & \cdots & \rho_{0T} \\ \rho_{10} & 1 & \cdots & \rho_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{T0} & \rho_{T1} & \cdots & 1 \end{bmatrix}$$

Where $\rho_{kl} = corr(Y_{ik}, Y_{il})$ for the i^{th} subject at times k and l. Note that the diagonal elements of the correlation matrix will be 1 (since they are the correlation of a time point with itself). The correlation matrix is also symmetric since the correlation of time point k with timepoint l is represented twice in the correlation matrix (once with k as the row and l as the column and again with k as the column and l as the row); i.e. $\rho_{T0} = \rho_{0T}$.

The independent structure assumes that there is no correlation between the errors across observation time-points. The AR(1) correlation structure assumes an exponentially decaying pattern of correlation according to the temporal distance between the repeated measurements (i.e. the correlation between baseline and early assessment time points is greater than the correlation between baseline and the final assessment time point) [299]. By contrast, the CS correlation structure assumes a constant correlation between two distinct measurements regardless of temporal distance (i.e. the correlation between baseline and the first assessment

time point is the same as the correlation between baseline and the last assessment time point). The unstructured covariance matrix assumes no temporal pattern to the correlation structure.

The sample size for the multi-follow-up design can then be calculated, using the following 6 pieces of information:

1) The type I error (α)
2) The power $(1-\beta)$
3) The variance of the post-baseline responses (σ^2)
4) A contrast vector appropriate for the hypothesis being examined (a), as presented in Table 7.3.1
5) A vector of the expected mean difference between treatment groups at each time point (d)
6) And the intra-subject correlation matrix for the repeated measurements (C)

In general, there are three repeated measures methods commonly used to analyse multi-follow-up repeated measures data:

1) Use both the pre-treatment assessment response and all post-treatment assessments in the response vector: **RMA**;
2) Use the patient specific change scores (each post-treatment assessment minus pre-treatment): **CSA**;
3) Use the baseline/pre-treatment scores as a covariate in the analysis of post-treatment measurements: **ANC**.

Below we present sample size formula for multi-follow-up designs for the RMA, CSA and ANC baseline analysis methods [300, 301].

7.3.2.3 RMA

The sample size for testing a difference in the rates of change between two treatment groups including the baseline in the response vector is given by [300]:

$$n = \frac{(z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2 2\sigma^2 a' C^{-1} a}{(a' C^{-1} d)^2} \qquad (7.3.12)$$

Where a is the vector of linearly increasing time coefficients for example, $a' = (-1,0,1)$ for a design in which $m = 3$, C^{-1} is the inverse of the intra-subject correlation matrix of the repeated measurements, and d is the vector of differences between treatment group means across assessment time points.

Rearrangement of equation (7.3.12) yields the power calculation for a fixed sample size using the baseline as part of the response vector [295]:

$$z_{1-\beta} = \frac{a'C^{-1}d}{\sigma}\sqrt{\frac{n}{2a'C^{-1}a}} - z_{1-\frac{\alpha}{2}} \tag{7.3.13}$$

7.3.2.4 ANC

The linear trend scores are not generally independent of baseline values like simple pre-post difference scores. Therefore, the ANCOVA can be used to correct for dependence of linear trend scores on baseline differences as in the case of simple pre-post difference scores.

The correlation between the baseline measure and a weighted combination of repeated measures can be written as [295, 300]:

$$\rho_b = \frac{\sum_{l=0}^{T} a_l \rho_{l0}}{\sum_{k=0}^{T}\sum_{l=0}^{T} a_k a_l \rho_{kl}} \tag{7.3.14}$$

Where ρ_{k0} is the k^{th} element in the first column of the correlation matrix between repeated measurements, with the diagonal elements $\rho_{kl} = 1$.

Within the context of GLS the correlation between baseline scores and the slope of repeated measures is estimated as [295]:

$$\rho_b = \frac{a'a_{(1)}}{\sqrt{a'C^{-1}a}} \tag{7.3.15}$$

Where $a_{(1)} = [1,0,\ldots,0]$, a vector of length m with the first entry (for baseline) equal to 1 and all remaining elements equal to zero.

Therefore, the sample size per group for testing the difference in linear trends with baseline covariate between two groups is [295]:

$$n = \frac{(z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2 2(1-\rho_c^2)\sigma^2 a'C^{-1}a}{(a'C^{-1}d)^2} \tag{7.3.16}$$

This can be rearranged to isolate for power as:

$$z_{1-\beta} = \frac{a'C^{-1}d}{\sigma}\sqrt{\frac{n}{2(1-\rho_c^2)a'C^{-1}a}} - z_{1-\frac{\alpha}{2}} \tag{7.3.17}$$

7.3.2.5 CSA

Currently within the literature is no generic formulate developed for calculation of sample size for a multi-follow-up change score analysis. This likely relates to the fact that calculation of change scores potentially alters the correlation strength and structure, resulting in a discordance between the correlation matrix of raw scores and the correlation matrix of change scores.

Let y_0 represent the baseline measurement, and y_k and y_l represent the responses at two post baseline measurements with $k \neq l$.

The correlation between the two change from baseline scores $y_k - y_0$ and $y_l - y_0$ is given by [302]:

$$Corr(y_l - y_0, y_k - y_0) = \frac{\rho_{y_k y_l}\sigma_{y_k}\sigma_{y_l} - \rho_{y_k y_0}\sigma_{y_k}\sigma_{y_0} - \rho_{y_l y_0}\sigma_{y_l}\sigma_{y_0} + \sigma_{y_0}^2}{\sqrt{\sigma_{y_0}^2 + \sigma_{y_l}^2 - 2\rho_{y_0 y_l}\sigma_{y_0}\sigma_{y_l}}\sqrt{\sigma_{y_0}^2 + \sigma_{y_k}^2 - 2\rho_{y_0 y_k}\sigma_{y_0}\sigma_{y_k}}} \tag{7.3.18}$$

To illustrate the disruption of the correlation matrix let us assume a hypothetical trial in which $m = 3$, in which the correlation structure between the 3 response time points (y_0, y_1 and y_2) is compound symmetric:

$$C = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

The discordance between raw and change scores is easily illustrated if we assume that there is zero correlation between y_0, y_1 and y_2 and to simplify let us assume that $\sigma_{y_0} = \sigma_{y_k} = \sigma_{y_l} = 1$

$$Corr(y_l - y_0, y_k - y_0) = \rho_{l-0,k-0} = \frac{\rho_{kl} - \rho_{k0} - \rho_{l0} + 1}{\sqrt{2 - 2\rho_{l0}}\sqrt{2 - 2\rho_{k0}}} = \frac{1-\rho}{2-2\rho} = \frac{1(1-\rho)}{2(1-\rho)} = \frac{1}{2}$$

This simple example shows that for any compound symmetry correlation structure (including the absence of correlation between response time points), the correlation of the change scores will be equal to 0.5. This specific correlation strength and structure is harder to predict when the correlation structure of the raw scores is complex (i.e. autoregressive or unstructured). Equation (7.3.18) reveals that the correlation matrix of change scores will not match the correlation of raw scores and this added complexity is likely the reason that a generic sample size formula for multi-follow-up using change scores has not been developed.

7.4 Discussion

Determining the sample size in repeated measures designs is complicated by the need to estimate the correlation of repeated measurements taken from the same subject. Furthermore, the use of the baseline measurement in the analysis of repeated measures data further complicates sample size estimation process. This review illustrates different uses of the baseline measurement in computing power for both pre-post and multi-follow-up repeated measures experimental designs.

The formulae presented show a clear distinction between methods of sample size for pre-post designs and multi-follow-up designs. When there are only two measurements, a pre-post design, change score analysis is less efficient than analysing only the post-treatment measurements if the correlation between the baseline score and endpoint score is less than 0.5 [303]. Using the baseline as a covariate in the analysis of post-treatment measurements is more efficient than either the change scores, or post-treatment only analyses across all values of correlation are greater than zero.

For the analysis of multi-follow-up data, it is not clear if the use of change scores, or baseline as a covariate offers an efficiency gain over retaining the baseline measurement as part of the response vector. This is because 1) removal of the baseline response, either by inclusion as a covariate (ANC) and or through the calculation of change scores (CSA) will reduce the dimensions of the correlation matrix as compared with analyses that keep the baseline as a response; and 2) using change scores will alter the correlation structure as compared with the raw scores. The impact of these two features will likely depend on the correlation structure of the raw scores. Previous researchers have demonstrated that the required sample size always decreases as the

number of measurements per subject increases under the compound symmetry (CS) correlation [299]. By contrast, making additional measurements from each subject may increase required sample size under AR(1) correlation structure [299, 301, 304].

In the next chapter, I will assess the relative efficiency of these baseline analysis approaches needs for various correlation structures within the multi-follow-up setting.

Chapter 8. Sample size calculations for repeated measures longitudinal studies

8.1 Abstract

This chapter explores the basic factors that determine an appropriate sample size by contrasting methods for its calculation incorporating intra-patient correlation of response data. We found that the number of study assessment time points, level of intra-patient correlation and correlation structure all influence the required number of subjects to maintain a fixed type-I and type-II error rate. There is no reduction in sample size associated with moving from a study with a single follow-up time point (pre-post design) to a study with baseline and two follow-up measures; however further increases in the number of post-baseline patient assessments reduces the required number of subjects to achieve statistical power. Failure to account for the correlation and the use of sample size methods that focus only on the mean treatment difference at the end of the trial may result in a large over-estimate of the required number of subjects per treatment group and consequently may result in a gross misallocation of trial resources.

8.2 Introduction

Clinical trials often use parallel-group longitudinal designs in which patients are randomly assigned between treatment groups, evaluated at baseline, administered with the intervention(s) being examined and then evaluated again at intervals across a treatment/follow-up period of fixed duration [305]. Such trial designs yield multiple or 'repeated' measurements on each subject [297, 306]. There are two distinct advantages to longitudinal trials as compared with a cross sectional design: 1) control for baseline imbalances [307, 308] and 2) increased power for a fixed sample size [303, 309]. The hypothesis of primary interest is usually the difference between treatment groups in both the pattern and magnitude of change from baseline with respect to the primary endpoint. It follows that the analysis methods employed to the data collected from such a trial would aim to elucidate the difference between treatment groups with respect to the average rate of change measured from the slopes of the regression lines fitted to the mean response patterns. A critical step in planning such a clinical trial is determining the sample size that will detect an effect of a given magnitude, or to estimate the power with which effect of particular magnitude

can be detected given a fixed sample size [310, 311]. A sample size insufficient for the effect size will result in reduced power to detect a clinically meaningful difference. Conversely, an inflated sample size potentially exposes patients to inferior treatments. Both insufficient and inflated sample sizes represent trial inefficiency and a waste of resources, for this reason accurate estimation of required sample size remains paramount in trial design [312, 313].

When the outcome variable is measured only once for each experimental unit (subject), methods for determining power and sample size are well-established and relatively straight forward to calculate [314]. By contrast, when a repeated measures experimental design is employed the estimation of the appropriate sample size is complicated by estimation of the structure and strength of the within patient correlation of repeated measurements [292]. Typically, there is little to no information on these additional parameters at the time of study planning [315] and the assumptions used in the sample size calculation are little more than guesses. The consequences of under- or over-estimating the intra-patient correlation of responses are complicated, and will likely depend on both the effect size and number of repeated measurements.

In this chapter, we contrast methods for calculating sample size for studies with repeated measurements of normally distributed continuous responses. We consider a range of post baseline assessment time points and intra-patient response correlation and two alternate correlation structures. In addition to examining the specific sample size requirements we consider the implications of incorrectly assuming a given correlation structure. The methods and results have direct implications for the design of clinical trials.

8.3 Methods

8.3.1 Notation

We will continue with the notation outlined in section 7.2.1. Additionally, we use D_k to denote the difference between group means at time point k, i.e D_T is the difference between group means at the final assessment time point:

$$(D_T = \mu_{1T} - \mu_{0T})$$

Increasing the number of repeated measurements within a trial can result in following patients for a longer time or more frequent assessment within a fixed observation (**T**). The latter is the

perspective used throughout our work: that is, in moving from $m = 2 \rightarrow m = 3 \rightarrow \cdots \rightarrow m = \infty$, both the difference between groups at the final assessment time point ($\mu_{1T} - \mu_{0T} = D_T$) and the correlation between baseline and the final assessment time point ($\rho_{0,T}$) remains equal irrespective of the number of patient assessments (Figure 8.3.1). This has two important consequences: firstly, the interval of time between assessment time points must decrease with increasing m or the slope of each treatment arm would decrease, and secondly, we cannot explore baseline imbalance in these calculations and keep D_T constant.

Lastly, we define the slopes of each treatment arm as:

$$\begin{aligned} Control: Slope_0 &= \frac{\mu_{0T} - \mu_{00}}{T} \\ Active: Slope_1 &= \frac{\mu_{1T} - \mu_{10}}{T} \end{aligned} \tag{8.3.1}$$

And the difference in slopes between treatment arms is: $Slope_1 - Slope_0 = \Delta$. For the reasons mentioned previously, we fix the intercept at 0 for both treatment arms:

$$Slope_1 - Slope_1 = \frac{\mu_{1T} - \mu_{10}}{T} - \frac{\mu_{1T} - \mu_{10}}{T} = \frac{\mu_{1T}}{T} - \frac{\mu_{0T}}{T} \tag{8.3.2}$$

Hence:

$$Slope_1 - Slope_0 = \mu_{1T} - \mu_{0T} = D_T = \Delta \tag{8.3.3}$$

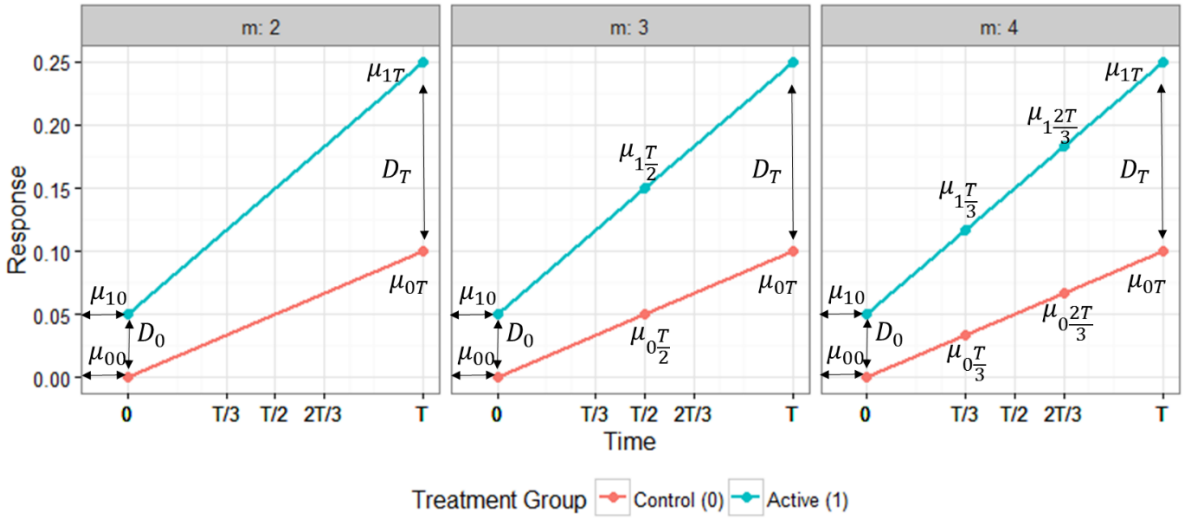


Figure 8.3.1 Hypothetical trial scenarios.
m:2 has only has one follow-up measurement, m:3 and m:4 have two and three follow-up measurements respectively. Each panel has the same initial difference (D_0), final difference (D_T) for the treatment groups. The treatment specific slopes (S_j) are calculated from $S_j = (\mu_{jT} - \mu_{j0})/T$. Consequently, by reducing the interval between assessments, more frequent follow-up, the treatment specific slopes remain constant across panels.

8.3.2 Sample size equations

8.3.2.1 Two sample t-test

The equation for estimation of the sample size required for adequate power in testing the significance of difference between means is given by [294, 313]:

$$n = \frac{2\sigma^2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{{D_T}^2} \qquad (8.3.4)$$

Where $\sigma^2$is the common variance in the comparator groups, $z_{1-\alpha/2}$ is percentile of the standard normal distribution corresponding to the desired alpha level (the Z score beyond which the smaller area under the normal curve is equal to α for a one-sided test or α/2 for a two-sided test) and $z_{1-\beta}$ is percentile of the standard normal distribution corresponding to the desired power. Hereafter we refer to Equation (8.3.4) as the **TT method** (derived from its use in calculating the sample size used in t-tests).

8.3.2.2 Repeated measures sample size

In the case of repeated measurements, the equation for estimation of the sample size required for adequate power in testing the significance of difference between means is given by [300, 313]:

$$n = \frac{2(z_\alpha + z_\beta)^2}{\delta^2} \quad (8.3.5)$$

Where:

$$\delta = \frac{\varphi_c}{\sigma_c} \quad (8.3.6)$$

The numerator (φ_c) is a quantity describing the contrast weighted difference in means between comparator groups at a particular time point:

$$\varphi_c = \sum_{k=0}^{T} c_k(\mu_{1k} - \mu_{0k}) \quad (8.3.7)$$

The denominator (σ_c) is a combination of the sum of contrast weighted common variance at each time point and the sum of each distinct time point pair combination (i.e. time points k and l, where k≠l) contrast coefficients multiplied by the covariance between time points k and l:

$$\sigma_c^2 = \sum_{k=0}^{T} c_k^2 \sigma_l^2 + 2 \sum_{k<l}^{T} c_k \, c_l \sigma_{kl} \quad (8.3.8)$$

Hereafter we refer to equations (8.3.5) as the **OD method**, abbreviated from the authors initials.

8.3.2.2.1 Contrast coefficients

A repeated measures design examining only the effect of treatment over time results in one repeated 'within-subject' fixed factor – time, one 'between-subjects' factor – treatment and the cross of the 'within-subject' and 'between-subject' factors. This yields three separate null hypotheses relating to each factor and the cross of factors. The general omnibus test from an ANOVA only tells you if one or more of these null hypotheses has been rejected, but does not inform you of which specific hypotheses. Orthogonal contrasts allow the total sum of squares (SST) to be partitioned (i.e. decompose from the uninformative H_0 of the omnibus ANOVA) into meaningful and targeted comparisons of the treatment means at each time point, or one for each contrast:

$$SS_{Total} = SS_{contrast1} + SS_{contrast2} \ldots + SS_{contrasti} \quad (8.3.9)$$

This partitioning of SST among its degrees of freedom is a consequence of the orthogonality of the contrasts. The contrast coefficients are obtained by orthogonal polynomial coefficients to the repeated measurements. Table 7.3.1 in the previous section presents the linear coefficients for equally spaced assessment time point. Of note, the sum of all contrast coefficients squared is always equal to the negative of twice the sum of each contrast coefficient multiplied by the preceding coefficients such that:

$$\frac{\sum_{k<l} c_k c_l}{\sum {c_k}^2} = -\frac{1}{2} \qquad (8.3.10)$$

8.3.2.3 Covariance structures

The covariance between the responses between two time points - σ_{kl} is a function of the variance and correlation. Often there is a trend or pattern in the correlation between response time points which yields a variance/covariance structure (hereafter called simply the covariance structure**).** Two common covariance structures often encountered in medical data are compound symmetry (CS) and autoregressive (AR(1) – hereafter referred to as AR). CS correlation structure assumes a uniform correlation positive correlation between all assessment time points while the AR correlation structure assumes that the correlation between responses on the same patient decay towards zero as the time separation between the responses increases (see section 0).

8.3.3 Parameters examined

In repeated measures linear regression, the variable of interest is usually the treatment difference in the change from baseline – i.e. the difference in slopes between treatment arms (Δ).

Table 8.3.1 Parameters used in sample size calculations

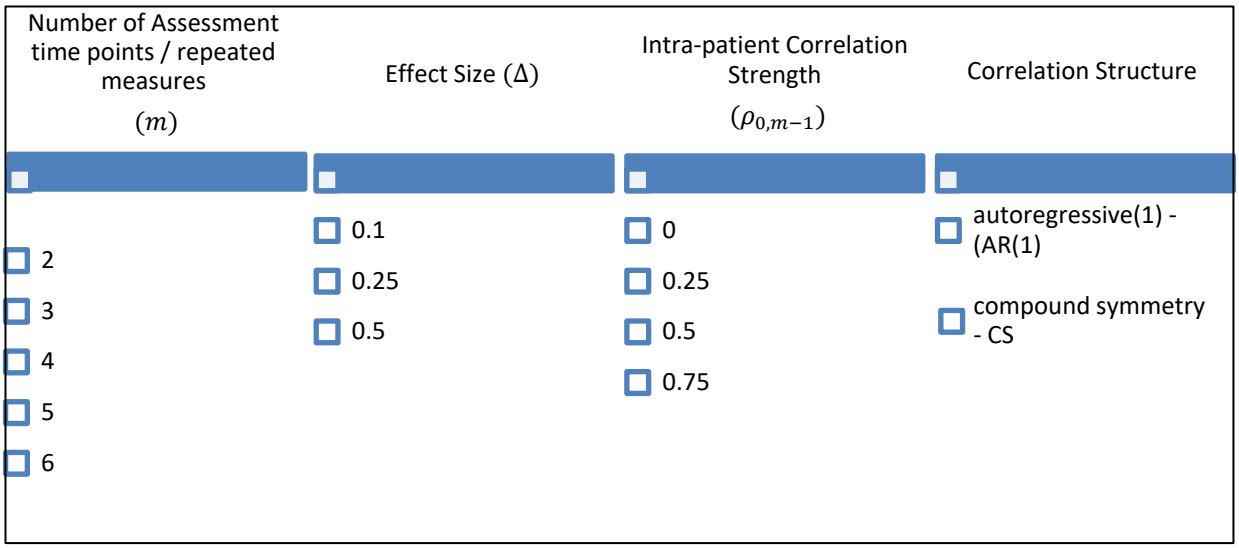


Table 8.3.1 lists the values of parameters used in each of the sample size methods. We considered three alternate values for the effect size (Δ): 0.1, 0.25 and 0.5. The correlation coefficient ranged between 0 and 0.75 in increments of 0.25. Five different number of assessment time points (m) were considered: 2- 6.

Therefore, design parameters covered 120 different situations: 5 (number of repeated measures) x 3 (values of the slope difference) × 4 (values of correlation strength) ×2 (correlation structures).

For each of the 120 different trial scenarios we calculated the required sample size per group using the TT method (8.3.4) and again using the OD method (8.3.5). For both sample size calculations, the value of the intercept was fixed at 0 and the variance (σ^2) was chosen to be 1 at all m time points. Consequently, each examined effect size can be thought of as a Cohen's-D effect size. Alpha was set at 0.05 and beta at 0.2, corresponding to conventional design parameters of 5% Type-I error and 80% power.

8.4 Results

8.4.1 Sample size estimation

Table 8.4.1 shows the required sample size given various numbers of assessments (m) effect sizes (Δ), correlation strengths $(\rho_{0,T})$ and correlation structures (AR(1) and CS) assuming $\beta = 0.2$ (80% power) and $\alpha = 0.05$ (5% Type-I error). The 'TT method' column provides the sample size calculation using (8.3.4), and is based only on the difference between treatment group means at

the final assessment time point $(m-1)$. Consequently, for a fixed α and power, the estimated sample size (n) is insensitive to changes in the number of assessment time points (m), intra-patient correlation strength $(\rho_{0,m-1})$ and correlation structure. The autoregressive(1) columns display the required sample size calculated using equation (8.3.5) across alternate correlation strengths assuming an AR(1) correlation structure. Similarly, the compound symmetry columns display the required sample size per group calculated using equation (8.3.5) across alternate correlation strengths assuming a CS structure.

Table 8.4.1 Sample sizes for various combinations of effect size (Δ) and assessment time points.
Correlations between the baseline and final assessment time point are 0, 0.25, and 0.75. Results were calculated assuming both an AR(1) and CS correlation structure. The TT method refers to the required sample size calculated using only the difference between group treatment means at the final patient assessment while the OD method refers to sample size calculation incorporating both the number of repeated measures, correlation strength and structure

Number of assessment time points (m)	**Δ**	**OD Method**								**TT Method**
		Autoregressive(1)				**Compound Symmetry**				
		Correlation Strength $(\rho_{0,T})$								
		0	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**	
2	0.1	1570	1178	785	393	1570	1178	785	393	1570
	0.25	252	189	126	63	252	189	126	63	252
	0.5	63	48	32	16	63	48	32	16	63
3	0.1	1570	1178	785	393	1570	1178	785	393	1570
	0.25	252	189	126	63	252	189	126	63	252
	0.5	63	48	32	16	63	48	32	16	63
4	0.1	1413	1204	804	402	1413	1060	707	354	1570
	0.25	227	193	129	65	227	170	114	57	252
	0.5	57	49	33	17	57	43	29	15	63
5	0.1	1256	1235	824	411	1256	942	628	314	1570
	0.25	201	198	132	66	201	151	101	51	252
	0.5	51	50	33	17	51	38	26	13	63
6	0.1	1122	1263	841	418	1122	841	561	281	1570
	0.25	180	202	135	67	180	135	90	45	252
	0.5	45	51	34	17	45	34	23	12	63

Figure 8.4.1 provides a graphical representation of the information contained within Table 8.4.1. Both the table and figure show that larger effect sizes require smaller sample sizes across all three sample size calculation methods. When the correlation structure is compound symmetry (CS) the reduction is sample size is directly proportional to the increase in correlation. Using the derivation of σ_c^2 from (8.3.8), when the correlation is equal to zero $(\sigma_{ij} = 0)$ then:

$$\sigma_c^2 = \frac{1}{2}\sum_{j=1}^{m} {c_k}^2$$

If the correlation is increased such that $(\sigma_{kl} > 0)$ then the ratio of:

$$\frac{\sigma_c^2|\sigma_{kl} > 0}{\sigma_c^2|\sigma_{kl} = 0} = \frac{\frac{1}{2}\sum_{k=0}^{m-1} {c_k}^2 + \sum_{k<l}^{m-1} c_k c_l \sigma_{kl}}{\frac{1}{2}\sum_{k=0}^{m-1} {c_k}^2} = \frac{\frac{1}{2}\sum_{k=0}^{m-1} {c_k}^2}{\frac{1}{2}\sum_{k=0}^{m-1} {c_k}^2} + 2\sigma_{kl}\frac{\sum_{k<l}^{m-1} c_k c_l}{\sum_{k=0}^{m-1} {c_k}^2}$$

As previously noted in Equation (8.3.10):

$$\frac{\sum_{k<l} c_k c_l}{\sum {c_k}^2} = -\frac{1}{2}$$

$$\therefore \frac{\sigma_c^2|\sigma_{kl} > 0}{\sigma_c^2|\sigma_{kl} = 0} = 1 - 1\sigma_{kl}$$

This means that each percentage increase in correlation is matched with a percent decrease in sample size irrespective of the number of time points. When the correlation structure is autoregressive, the effect upon sample size is less obvious. Increasing the intra-patient correlation and number of repeated measures can reduce the required sample size, however, the effects of each are not consistent. We draw attention to two distinct features of Figure 8.4.1:

1) There is no difference in the estimated sample size from associated with increasing the number of repeated measures from 2 to 3 (m=2 →m=3)
2) For AR(1), data the change in sample size between ρ=0 →ρ=0.25 is always less than the reduction in sample size associated with ρ changing between 0.25 →1

To understand each of these phenomena we reformulate equation (8.3.5) as [295]:

$$n = \frac{2(z_\alpha + z_\beta)^2 \sum_{k=0}^{T} \sum_{l=0}^{T} c_k c_l \sigma_{kl}}{\left(\sum_{k=0}^{m-1} c_k(\mu_{0k} - \mu_{1k})\right)^2} \tag{8.4.1}$$

Thus, for a fixed z_α and z_β the equation presents the required sample size is a function of:

$$n \approx \frac{\sum_{k=0}^{T} \sum_{l=0}^{T} c_k c_l \sigma_{kl}}{\left(\sum_{k=0}^{T} c_k(\mu_{0k} - \mu_{1k})\right)^2} \tag{8.4.2}$$

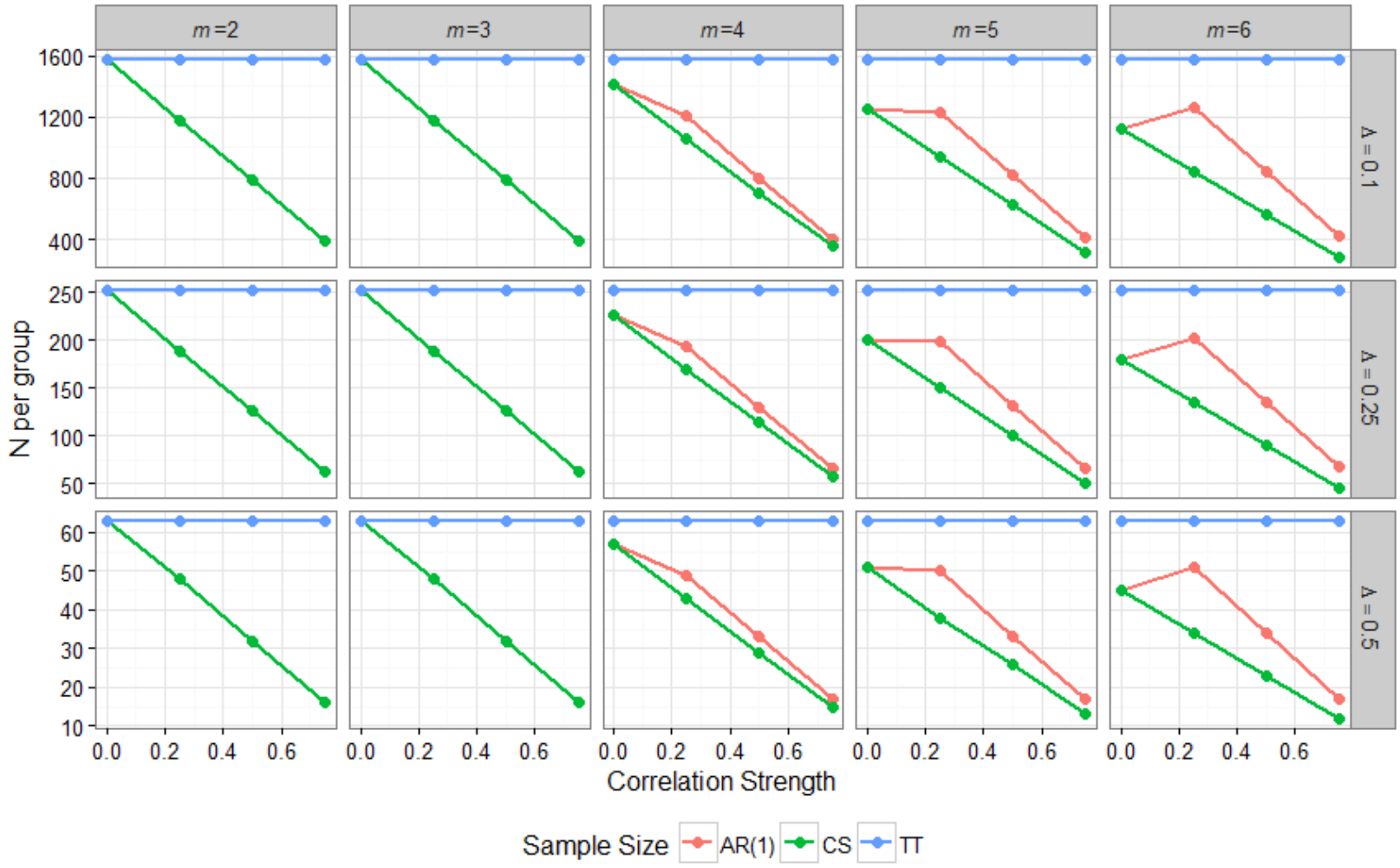


Figure 8.4.1 Relationship between sample size and correlation strength across alternate effect size and number of repeated measures.
TT presents the estimated sample size using the mean difference between treatment groups at the final assessment time point. Both AR(1) and CS use the repeated measures sample size calculation assuming AR(1) and CS correlation structures.

The numerator represents the multiplication of the contrast coefficients and covariance between alternate time points. The denominator is the contrast coefficient for a given time point multiplied by the treatment mean difference at that time point. Table 7.3.1 shows that the baseline ($k = 0$) and final ($k = T$) contrast coefficients are -1 and 1 respectively for both $m = 2$ and $m = 3$. For a fixed Δ, the group mean difference both at baseline and final assessment will be the same for $m = 2$ and $m = 3$, then these quantities will be the same. The addition of a 0 contrast coefficient will not add to either the numerator or the denominator of equation (8.3.7) and hence it stands to reason that the estimated sample size for $m = 2$ and $m = 3$ are equal.

The second phenomenon relates to the observed correlation in AR(1) structures. Table 8.4.2 shows the covariance matrix for AR(1) with 6 time points for $\rho_{0,m-1}$=0, $\rho_{0,m-1}$=0.25 and $\rho_{0,m-1}$=0.5.

Table 8.4.2 Covariance matrices for AR(1) for $\rho_{0,m-1}$ = 0, 0.25 and 0.5

$\boldsymbol{\rho = 0}$	$\boldsymbol{\rho = 0.25}$	$\boldsymbol{\rho = 0.5}$
$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.76 & 0.57 & 0.44 & 0.33 & 0.25 \\ 0.76 & 1 & 0.76 & 0.57 & 0.44 & 0.33 \\ 0.57 & 0.76 & 1 & 0.76 & 0.57 & 0.44 \\ 0.44 & 0.57 & 0.76 & 1 & 0.76 & 0.57 \\ 0.33 & 0.44 & 0.57 & 0.76 & 1 & 0.76 \\ 0.25 & 0.33 & 0.44 & 0.57 & 0.76 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.87 & 0.76 & 0.66 & 0.57 & 0.50 \\ 0.87 & 1 & 0.87 & 0.76 & 0.66 & 0.57 \\ 0.76 & 0.87 & 1 & 0.87 & 0.76 & 0.66 \\ 0.66 & 0.76 & 0.87 & 1 & 0.87 & 0.76 \\ 0.57 & 0.66 & 0.76 & 0.87 & 1 & 0.87 \\ 0.50 & 0.57 & 0.66 & 0.76 & 0.87 & 1 \end{bmatrix}$

For lower correlation values between the first and last time point, the disparity between the earlier and later time point correlation strengths is greater. As the earlier time points contribute more to the weighting when multiplied through the contrast coefficients then the $\sum_{k=0}^{T}\sum_{k=0}^{T} c_k c_l \sigma_{kl}$ quantity is increased at low correlation, which results in an inflation of the required sample size to achieve a given level of power. Further increases in the correlation strength between the first and last time point have a net effect of decreasing $\sum_{k=0}^{T}\sum_{k=0}^{T} c_k c_l \sigma_{kl}$ (Table 8.4.3) and consequently the required sample size to achieve a given level of power is decreased.

Table 8.4.3 Numerator from equation (8.3.7)

ρ	$\sum_{k=0}^{T}\sum_{l=0}^{T} c_k c_l \sigma_{kl}$	
	CS	**AR(1)**
0	70.00	70.00
0.25	52.50	78.81
0.5	35.00	52.47
0.75	17.50	26.08

8.4.2 Inflation of the sample size

Table 8.4.4 shows the ratio of the sample size calculated from the TT method relative to the sample size calculated using the repeated measures OD method. A value of one indicates that both the sample size based upon the final group mean difference (TT method) is the same as the estimated sample size using the repeated measures formula (OD method); a value of 2 indicates that the sample size calculated based upon the final difference in groups means is twice the required sample size achieve 80% power estimated from the repeated measure formula. Hence, values above 1 represent the inefficiency using the TT method to calculate the sample size. Under all examined conditions the estimated sample size using OD the method (8.3.5) is less than or equal to the required sample size estimated from the TT method (8.3.4). The inefficiency of using (8.3.4) to estimate the sample size is exacerbated by both intra-patient correlation and increasing the number of assessment time points. The TT method and OD method estimate similar sample size requirements when the intra-patient correlation is zero and there are less than 4 patient assessments. In an experiment with baseline and 5 follow-up assessments, if the true intra-patient correlation and correlation structure are 0.75 and compound symmetry, then the indicated sample size calculated based upon only the final group mean difference would be ~5 times the sample size required to provide 80% power.

Table 8.4.4 TT sample size relative to using repeated measures

Number of assessment time points	Δ	Autoregressive(1)				Compound Symmetry			
		Correlation Strength							
		0	0.25	0.5	0.75	0	0.25	0.5	0.75
2	0.1	1.00	1.33	2.00	3.99	1.00	1.33	2.00	3.99
	0.25	1.00	1.33	2.00	4.00	1.00	1.33	2.00	4.00
	0.5	1.00	1.31	1.97	3.94	1.00	1.31	1.97	3.94
3	0.1	1.00	1.33	2.00	3.99	1.00	1.33	2.00	3.99
	0.25	1.00	1.33	2.00	4.00	1.00	1.33	2.00	4.00
	0.5	1.00	1.31	1.97	3.94	1.00	1.31	1.97	3.94
4	0.1	1.11	1.30	1.95	3.91	1.11	1.48	2.22	4.44
	0.25	1.11	1.31	1.95	3.88	1.11	1.48	2.21	4.42
	0.5	1.11	1.29	1.91	3.71	1.11	1.47	2.17	4.20
5	0.1	1.25	1.27	1.91	3.82	1.25	1.67	2.50	5.00
	0.25	1.25	1.27	1.91	3.82	1.25	1.67	2.50	4.94
	0.5	1.24	1.26	1.91	3.71	1.24	1.66	2.42	4.85
6	0.1	1.40	1.24	1.87	3.76	1.40	1.87	2.80	5.59
	0.25	1.40	1.25	1.87	3.76	1.40	1.87	2.80	5.60
	0.5	1.40	1.24	1.85	3.71	1.40	1.85	2.74	5.25

8.4.3 The consequence of correlation structure misspecification

As previously presented in equation (7.3.13), the power for a fixed sample size can be obtained by solving the following formula:

$$z_{1-\beta} = \frac{a'C^{-1}d}{\sigma}\sqrt{\frac{n}{2a'C^{-1}a}} - z_{1-\frac{\alpha}{2}}$$

Where a is the vector of increasing contrast coefficients presented in Table 7.3.1, i.e. $a' = [-3 \quad -1 \quad 1 \quad 3]$ for $m = 4$, C^{-1} is the inverse of the intra-subject correlation matrix of repeated measurements and d is the vector of group mean differences at each time point i.e. $d = [0 \quad 0.083 \quad 0.167 \quad 0.25]$ for $\Delta = 0.25$.

Figure 8.4.2 presents the power calculations using the sample size estimated by correlation structure misspecification, that is the power that you would have if you estimated your sample size assuming CS structure when in actuality the true intra-patient correlation structure was AR(1) (and vice versa). These were calculated using the sample size estimates from Table 8.4.1 and equation (7.3.13). The full results of these calculations are presented in Appendix D.

The results show that for a fixed Δ and $\rho_{0,m-1} > 0$, the loss or gain in power association with misspecification is primarily a function of the number of assessment time points. Misspecification of AR(1) structure as CS at the time of sample size calculation results in a reduction of power proportional to the number of patient assessments. By contrast, misspecification of CS structure as AR(1) during the sample size estimation will result in increased power beyond 80%.

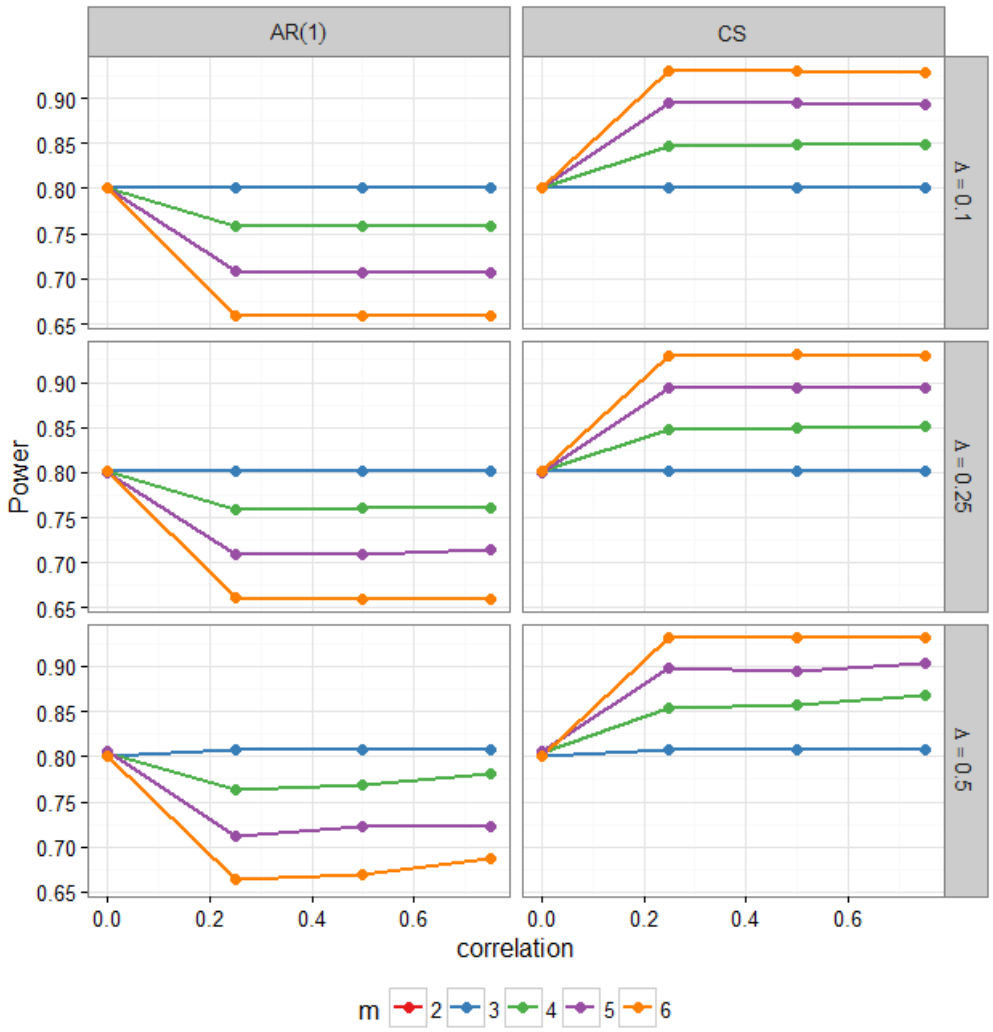


Figure 8.4.2 Power calculations under correlation structure misspecification.
The AR(1) column present the power calculation results using the sample size estimation assuming CS structure and the CS column presents the power calculation using the sample size estimation assuming AR(1) structure. Misspecification of the correlation structure results in a loss of power when there is an assumed CS correlation structure and an 'actual' AR(1) correlation structure. Conversely, misspecification results in a power gain when the assumed correlation structure is AR(1) and the true correlation structure is CS. The gain or loss in power is largely insensitive to correlation strength beyond 0 and is largely a function of the number of assessment time points.

8.5 Discussion

One of the most important aspects in clinical research design is the sample size estimation [295, 316]. Despite the wealth of literature regarding the need for trials to employ appropriate sample size calculations, sample size and statistical power remain relevant topics for biostatisticians today. Statistical inference of a trial relies upon an appropriate sample size in order to derive precise, accurate and robust conclusions. Presumably for this reason, the British Medical Journal and the regulatory authorities, including the FDA and the Committee for Proprietary Medicinal Products (CPMP 1995) in the EU, all require a justification on the planned study size [317]. Their reasons are captured in the guidelines of the international conference on harmonisation (ICH E9): essentially – if too few subjects are involved the study is potentially a misuse of time as realistic medical differences are unlikely to be distinguished from chance variation. Too large a study can be a waste of resources and in addition it is unethical to exposure patients to inferior medicine [318].

In the design of a randomised clinical trial with repeated measures of the outcome variable, the appropriate sample size calculation requires estimates of the variance of the outcome measure and the correlations among the measurements over time. In the absence of information regarding the intra-patient outcome correlation researchers frequently calculate the required sample size per group using methods appropriate for a two-sample t-test; a practice that is considered ‘ultra’ conservative by assuming that there is zero correlation between the baseline and follow-up measures [315, 319].

In any successful superiority trial, the patients in the control arm will have received an inferior treatment. If the patient group allocation is equal, this will be half of the study patient population. Even if the trial is unsuccessful (i.e. fails to establish superiority), then the experimental group has still been exposed to a new drug with potential unknown side effects. For these reasons, we contend that ultra-conservatism in the sample size is potentially unethical, and trialists calculating sample size should make every effort to minimise the number of exposed patients while maintaining power to detect clinically relevant effects. Ideally, assessments of the intra-patient correlation strength and structure will become routine in all early phases of clinical development and larger phase II and phase III studies will be able to exploit this information.

One important observation our results is that, under the AR(1) correlation structure, making additional measurements does not always reduce sample size requirement. Increasing the number of assessment measurements beyond m = 3 results in an increased sample size to maintain power when the correlation is intermediate-low (0.25). This counter-intuitive property has been observation by previous researchers who have attributed it to a theoretical property of the AR(1) model [299, 304]. This finding indicates that analysts should carefully consider the number of patient assessment in repeated measures trials where the response may is expected to exhibit AR(1) properties. While it may be tempting to conclude that for data with AR(1) structure and a moderate level of correlation, any design that has more than 2 measurements per patient would be sub-optimal, Zhang and Ahn [320] explored missing data with AR(1) models and concluded that increasing the number of measurements might be beneficial when the data suffer from missingness. A second important observation from this work is that increasing the number of assessment time-points from m=2 to m=3 has no impact upon the power of the study to detect treatment effect. This counterintuitive observation relates to the 2nd contrast coefficient of the linear trend being equal to zero. However, as the specific contrast coefficients are a function of both the spacing of measurement assessments as well as the expected response pattern, if the time points are not evenly spaced, or if the regression model was not limited to a linear trend then this contrast coefficient would have a non-zero value and the addition of the additional assessment measurement may be expected to increase power within the study.

Our results indicate that in the absence of information regarding the correlation structure and strength calculation of the required sample size assuming zero correlation and an AR(1) correlation structure can still take advantage of the repeated measures while still reducing the sample size when the number of post baseline assessments exceeds three. The choice of the specific number of repeated measures may reflect either a desire to characterise the time course of patient symptoms and/or to increase the precision of the estimate. If the former is true, then the number of repeated measures will and should depend on the investigators' expertise in patient safety monitoring. Alternatively, if the desire is to increase the precision of the estimate then our results indicate that there is no advantage in moving from a pre-post design to baseline and two follow-up assessments for a fixed sample size. This implies that the application of study

resources to improve precision might be better spent on more patients rather than increased monitoring if it is not possible to collect more than two post baseline responses.

This work has several important limitations that can be viewed as areas for future work and/or refinement. Firstly, we did not investigate empirically the increase in sample size associated with low correlation strength in the AR(1) structure. As sample size and power are directly linked, this result implies that low correlation with AR(1) will have less power than zero correlation. Future work could explore this through simulation. Secondly, we did not examine the influence of missing data. Patient attrition is an important factor when performing longitudinal studies. This topic has been studied previously in [313, 321-324]. Galbraith et al. [322] suggested computing the sample size requirement using 90% power when 80% power is intended thereby maintaining adequate power under moderate patient attrition. Similarly, Hedeker et al. [313] suggested inflating N by $1/(1 - f)$, where f is the greatest rate proportion of attrition at any of the assessment time points. Lastly, all the analysis models assumed a linear effect over time of treatment which simplifies the framework for the study but may be unrealistic. Extensions to this would be simulating a nonlinear effect over time, in particular, a treatment effect that decreases over time. The analysis could then use more flexible models to capture information that is lost when the linear model does not adequately describe the true underlying treatment effects in the data. Presumably, if we simulate with linear we would potentially lose power and precision. The magnitude of which could be investigated. Similarly, we have assumed equally spaced time points of assessment. Many trials may wish to monitor patients more frequently at the start of the observation period as compared with later during the trial. While it is possible to construct contrast coefficients for unequally spaced time points, our results are not generalisable to situations in which the assessment time points are not evenly spaced.

In conclusion, sample size calculation is a very important aspect of any study. It should be done at the time of planning a study, based on the type of the research question and study. More work is needed to understand the impact of study assumptions upon the required sample size and power. Furthermore, given the impact of intra-patient correlation upon power it should become standard for longitudinal studies to report details regarding the correlation parameters both at the time of study planning and retrospectively to see how the assumptions matched the observation.

Chapter 9. The analysis of longitudinal data

9.1 Abstract

Longitudinal experimental designs frequently collect multiple observations of an outcome across time on a set of subjects within a given study. The data from such studies are commonly referred to as 'repeated measures'. Frequently observations from the same subject are positively correlated. Consequently, statistical analysis of repeated measures data must account for this correlation. In this review, we explore two analysis methods commonly employed when analysing repeated measures data: 1) the marginal model, and 2) the mixed effect model. The methods are described within the context of linear regression however both techniques are also applicable to non-linear data scenarios. Extensions of the linear model to account for alternate baseline analysis strategies are also described.

9.2 Overview of analysis methods

Clinical trials frequently employ a repeated measurements design in which individuals are randomly assigned between two (or more) treatment groups, measured at baseline, and then measured at fixed intervals throughout a specified treatment period [300]. In such an experiment the primary goal is usually to explore the change in response over time and the factors that influence change [325].

A distinctive feature of repeated measures data is that they are 'clustered' – observations from the same patient (cluster) will typically exhibit positive correlation [326]. Thus, to analyse these data ignoring this correlation, would contradict the crucial assumption of independence that is the cornerstone of simple linear regression [74, 326-328]. Consequently, statistical models for repeated measures data must explicitly describe and account for this correlation. Several approaches have been proposed to quantify treatment group differences in outcomes that arise from repeated measures studies. Most of the approaches can be grouped into one of two classes: marginal models (also known as population-average models) and mixed models (also known as subject-specific models) [329].

We will briefly compare each approach for the methods in which they account for intra-patient correlation of measurements.

9.3 Marginal model

Marginal or population-averaged models represent an extension of generalised linear models [330] to longitudinal data in which the within-subject correlation among the repeated responses from the same individual are incorporated into the model [331]. Marginal models describe how the mean response in the population changes over time and how these changes are related to fixed effect covariates [327]. Within this context, the term marginal refers to the fact that the model for the mean response at each occasion depends only on fixed effect covariates, and not on any random effects. This contrasts with mixed-effects models, where the mean response depends not only on fixed effects but also on subject specific random effects (this will be discussed in further in 9.4). A key aspect of marginal models is that the mean response and within-subject association are modelled separately. Consequently, the accuracy of model estimates is not altered in any way by the assumptions made about the structure or magnitude of the within-subject association and it is the precision of model estimates that incorporates the within-subject correlation.

It is difficult to pinpoint the origin of marginal models within the literature [327]. In the case of linear models, both the repeated-measures-ANOVA [332] and MANOVA [333] analysis methods fit within the framework of marginal models [327]. Like marginal models, repeated-measures-ANOVAs are also capable of modelling intra-subject correlation, however such models assume that the covariance structure is compound symmetry and that the variance is constant across time [327, 328]. While the MANOVA model allows for a flexible variance-covariance structure for the repeated measures and non-constant variance, it requires complete data for all subjects and identical measurement occasions [328] . Consequently, marginal models represent a more flexible form of repeated-measures-ANOVA and/or MANOVA where complex covariance structures, non-constant variance, and non-identical measurement times are permitted.

9.3.1 Hypothetical trial example

To illustrate how marginal models work, consider a hypothetical parallel group randomised control trial (RCT) of two treatments, where the primary response variable (Y) is a continuous variable. The primary response variable Y_{ik} (for ith subject, at the kth measurement, $i = 1, \dots, N$; $k = 0, \dots, T$) is scheduled to be observed at regular intervals for assessing change from baseline. In this scenario, we consider the common case of repeated measures RCT design in which the primary response variable is measured once during the baseline period (prior to randomisation), and T times dusting the treatment period (post randomisation), hence $k = 0$ indexes the pre-randomisation baseline measurement and T is the number of post-randomisation observations for each subject which combined create a $T + 1$ dimension vector of response for subject i:

$$Y_i = (\underbrace{Y_{i0}}_{\text{baseline data}}, \underbrace{Y_{i1}, \dots, Y_{iT}}_{\text{post randomisation data}})'$$

Assuming an average linear trend for Y as a function of time, a multivariate regression model of the treatment effect upon change from baseline can be obtained by assuming the elements Y_{ik} in Y_i satisfy:

$$Y_{ik} = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik}$$

With the assumption that the error components ε_{ik} are normally distributed with mean zero.

As a departure from the notation in the previous section, in this model, x_{1i} denotes the subject treatment group assignation, where:

$$x_{1i} = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ subject is assigned to the treatment group} \\ 0, & \text{if } i^{th} \text{ the subject is assigned to the control group} \end{cases}$$

This change in notation allows us to drop *j* from the response and predictor subscript and aids in the clarity of the presentation.

x_{2k} represents the time point of the observation where:

$$x_{2k} = \begin{cases} k, & \text{for k } > 0 \ (after\ randomisation) \\ 0, & \text{for k} = 0 \ (\text{before randomisation}) \end{cases}$$

In vector notation, we get:

$$Y_i = X_i\beta + \varepsilon_i$$

A design matrix X_i incorporating an intercept, treatment, time and treatment by time interaction would result in $\beta' = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\varepsilon_i' = (\varepsilon_{i0}, \varepsilon_{i1}, \dots, \varepsilon_{iT})$

The model is completed by specifying an appropriate covariance matrix V_i for ε_i, which estimates the variance-covariance of observations within Y_i, leading to the multivariate model [74, 334]:

$$Y_i \sim N(X_i\beta, V_i)$$

If $V_i = \sigma^2 I_T$, where I_T is an identity matrix of dimension T, then the model above would correspond to a simple linear regression model in which all Y_{ijk} observations are assumed to be independent, i.e. ignoring the potential for observations within a subject to be correlated. (Specification of the R matrix relaxes the assumption that within-subject errors are independent.)

The marginal model replaces the identity matrix of the residuals with a parameterized working correlation matrix R, such that $V_i = R_i$. To address correlated data, the working correlation matrix imposes structural constraints and it is necessary to specify the structure of correlation before it is possible to calculate the estimated correlation strength between time points. The choice of correlation structure depends on both prior knowledge of the endpoint as well as the observed data [74, 335]. The simplest model is the independent covariance model, where the within-subject error correlation is zero, and hence $V_i = \sigma^2 I_T$. The most complex is the unstructured covariance model, where within-subject errors for each pair of times have their own unique correlation [74]:

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho_{01} & \rho_{02} & \cdots & \rho_{0T} \\ & 1 & \rho_{12} & \cdots & \rho_{1T} \\ & & 1 & \cdots & \rho_{2T} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}$$

Where $\rho_{kl} = corr(Y_{ik}, Y_{il})$ = for the i^{th} subject at times k and l.

It is important to select an appropriate covariance model to infer accurate conclusions from repeated measures data analysis [336]. Failure to account for correlation by using a model that is too simple can potentially increase the Type I error rate by underestimating standard errors of the model coefficients [336]. Choosing a covariance structure that is too complex reduces both the power and efficiency of the analysis [337]. Both Diggle [338] and Wolfinger [335] recommend that analysts first explore the data using more complex (i.e. UN or AR) covariance structures. This helps avoid misspecification biases that can occur with models that are too simple. One common strategy employed by trialists for selecting the correct covariance structure is to fit a few possible candidate structures and then compare them using various information criteria [336].

9.4 Mixed models

Mixed-effect regression models (MRMs) are a popular choice for the analysis of longitudinal data. The underlying premise of linear mixed models is that some subset of the regression parameters varies randomly from one individual to another. Within this context, factor effects are considered random if they are used in the study to represent only a sample of a larger set of potential levels. In a mixed model, the mean response is modelled as a combination of fixed effects (factors which are shared by all individuals in the study i.e. treatment, time and their interactions) and subject specific random effects that are unique to an individual.

Like marginal models, MRMs do not have restrictive assumptions concerning assessment times across patients and the variance-covariance structure of the repeated measures [339, 340]. MRM's are considered to be robust to missing data and can easily account for time-varying covariates [339]. However, the primary reason why some researchers favour these models over marginal models is their ability to estimate change for each subject, hence their alternate moniker, 'subject-specific models'. The ability to estimate individual change across time is particularly useful in identifying subjects that deviate from the average treatment group trend [339]. Since their introduction by Laird and Ware [75], variants of MRM's have been developed under a variety of names: random coefficient models [341], mixed models [335, 342]. random regression models [340] models [340].

Continuing our hypothetical trial example from the previous section let us consider a mixed model in which each subject is considered to have a random intercept and slope. Assuming an average linear trend for Y as a function of time, a multivariate regression model of the treatment effect upon change from baseline can be obtained by assuming the elements Y_{ik} in Y_i satisfy:

$$Y_{ik} = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_1 x_{2k} + \beta_3 x_{1i} x_{2k}}_{\text{Fixed Effects}} + \underbrace{\tau_{0i} + \tau_{1i} x_{2k} + \varepsilon_{ik}}_{\text{Random Effects}}$$

Where:

$\tau_{0i} = \{\text{is the intercept deviation for subject } i, \text{ and}$

$\tau_{1i} = \{\text{is the slope deviation for subject } i$

If individuals did not deviate at baseline or in slope, then the τ_{0i} and τ_{1i} terms would equal 0. However, it is more likely that subjects will have positive or negative deviations from the population average intercept and/or slope and hence the τ_{0i} and τ_{1i} terms will deviate from 0. The most common form for the distribution of these terms is the normal distribution with mean 0 and variances $\sigma_{\tau 0}$ and $\sigma_{\tau 1}$. The variance term $\sigma_{\tau 0}$ represents the spread of the subject intercept terms while the $\sigma_{\tau 1}$ variance terms represents the spread of the subject slopes over time.

This model assumes that the errors of measurement are conditionally independent, which is to say that errors of the response observations are independent conditional on the random individual-specific effects of τ_{0i} and τ_{1i}.

Let q denote the number of random effects in the model, as this model has two random effects (intercept and slope), $q = 2$. In vector notation, we get:

$$Y_i = X_i\beta + Z_i u_i + \varepsilon_i$$

Where β denotes fixed effects with design matrix X_i (like the marginal model), u_i represents the $q \times 1$ vector of random effects with design matrix Z_i of dimensions $T \times q$, and ε_i is the $T \times 1$ vector of random error.

Both the random effects (u_i) and error ε_i are expected to be normally distributed with mean 0, and variances D and R respectively, that is $u_i \sim N(0, D)$, and $\varepsilon_i \sim N(0, R_i)$. While both the D and R_i

matrices are covariance matrices they have different dimensions. The R_i matrix is of dimension $T \times T$, while the D variance-covariance matrix is of dimension $q \times q$, i.e. if there were only one random effect per subject then D would be a 1×1 matrix, and for our example where we have both a random intercept and slope, D would be a 2×2 matrix.

Again, the mixed model is completed by specifying an appropriate covariance matrix V_i,

$$Y_i \sim N(X_i\beta, V_i)$$

only now the V_i matrix must account for the covariance between random effects as well as the covariance between subject errors. Hence the multivariate model can be written as:

$$Y_i \sim N(X_i\beta, ZDZ' + R_i)$$

Like the R_i matrix from the marginal model, a covariance structure must be specified for the variance-covariance matrix for the random effects (D). Typically, when the D matrix is used to specify the covariance of the Y_{ik} repeated measures, the structure of R_i is reduced $\sigma^2 I_T$ however it is possible to model the variance-covariance of observations within Y_i through both the D and R_i matrices simultaneously within the mixed model framework.

9.5 How to choose between a marginal and mixed model

Lui et al. [343] examined the use of D, R or both matrices simultaneously to model intra-patient correlation under three commonly used covariance structures - unstructured (UN), compound symmetry (CS), and auto-regressive (AR(1)). Based upon the results of their simulations, the authors concluded that using only the R matrix (marginal model) is recommended for data with either unstructured or compound symmetry covariance structure. For autoregressive data, if the variance of the random between subject effects is significant, then using both the D and R matrices results in accurate estimation of model parameters however using a marginal model rather than a mixed model with both R and D matrices specified had similar Type I and II error rates. Similarly, other authors have found that when the within-subject error covariance matrix specification (R) was combined with a D matrix of random intercepts and slopes models often do not converge; this has been attributed to over specification, since large numbers of sometimes

redundant covariance elements can result when both the D and R matrices are used within a model [337, 344, 345].

9.6 Baseline analysis strategies

One common feature of repeated measures RCTs that complicates the analysis is the presence of a baseline measurement. Baseline values of the outcome variable collected prior to randomisation and treatment administration are almost ubiquitously measured in clinical trials. While there are several reasons why the baseline measurement might be collected, the primary reason is to ensure that treatment groups are comparable prior to the administration of the intervention under examination. If the baseline outcome values are similar between treatment arms this reassures reviewers that the randomisation was successful and that differences in the outcome following treatment administration likely relate to the differences in treatment rather than the populations contained within each treatment group. Despite the reason for the baseline collection, statisticians are typically reluctant to test the equality of baseline values between treatment groups. Such a test, if performed, appears topically to represent a test of the randomisation process. However, any test will have a type I error associated with its performance and it may be expected that if a difference between treatment groups is observed it may not reflect a failing of the randomisation. Instead statisticians typically have sought ways to assess the treatment effect once removing any observed baseline imbalance. Consequently, the question arises as to how to handle the baseline measurement in the measurement in the assessment the patterns of change in the mean response over time are the same in the treatment groups.

Fitzmaurice et al. [326] outlines four common strategies used to incorporate the baseline value in repeated measures analysis:

1) Retain it as part of the outcome vector and make no assumptions about group differences in the mean response at baseline (RMA).
2) Retain it as part of the outcome vector and assume the group means are equal at baseline (cLDA).
3) Subtract the baseline response from all the remaining post-baseline responses and analyse the change from baseline (CSA).
4) Use the baseline value as a covariate in the analysis of the post-baseline responses (ANC).

We will consider each method in turn continuing with the hypothetical trial scenario outlined in section 9.3.1, a situation in which we have randomised subjects into two groups, a treatment and a control group.

A regression model for the RMA is given by:

$$Y_{ik} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik}$$

Where Y_{ik} is the response score for subject i $(i = 1, \dots, n)$ at time $k = (0, \dots, \mathrm{T})$. x_{2k} denotes the time point of assessment at $k = (0, \dots, \mathrm{T})$, and ε_{ik} is the patient by time error vector.

The RMA has also been called the longitudinal data analysis (LDA) [325], however as previously noted, with the absence of missing data and identical observation periods, this model is identical to a repeated-measures-ANOVA and hence we have used the abbreviation RMA to identify it throughout our work.

The cLDA (constrained longitudinal data analysis) model was proposed by Liang and Zeger [325] represents a modification of the RMA in which the baseline values are assumed to be equal in the intervention groups ($\beta_0{}^*$). Utilising the same notation as above, the regression model for the cLDA is given by:

$$Y_{ik} = \beta_0{}^* + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik}$$

Both these first two strategies retain the baseline measurement as part of the outcome vector but differ in the assumptions about the mean response at baseline. The first strategy corresponds to a standard analysis of the response profiles using all available data without any constraints on the group means at baseline [326]. This is the method of analysis introduced in sections 9.3 and 9.4. The second strategy corresponds to an analysis of response profiles using all available data where the treatment group means at baseline are constrained to be equal. In randomised controlled trials (RCTs), it may be expected that random assignment of patients to treatment groups will result in baseline balance and thus both methods may be expected to yield similar results. By contrast, in observation trials where there is no reason to assume that that comparator groups will be similar at baseline the cLDA strategy may not be appropriate [326].

The third strategy (CSA) uses change from baseline scores as the response variable rather than the raw response observations. A regression model for the change score is given by:

$$d_{ik}|\mathrm{k} > 0 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik}$$

Where:

$$d_i = (Y_{i1} - Y_{i0}, Y_{i2} - Y_{i0} \dots, Y_{iT} - Y_{i0})$$

The fourth strategy (ANC) analyses the post-baseline response measurements and incorporates the baseline response by including it as a covariate. A regression model for the ANCOVA is given by:

$$Y_{ijk}|\mathrm{k} > 0 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \beta_4 Y_{i0} + \varepsilon_{ik}$$

Where Y_{i0} represents the baseline response measurement for subject i.

Both the third (CSA) and fourth (ANC) analysis strategies do not use the baseline measurement as part of the response vector. This has important consequences for experiments in which $T = 1$, and is discussed further in section 9.7.1.

9.6.1 A fifth analysis option – analysis of covariance of change scores

Several authors have pointed out that there is no difference between an analysis of change scores using the baseline as covariate and an analysis of raw scores using baseline as covariate. As both response vectors differ by Y_{i0}, and both analyses estimate model effects that are adjusted for Y_{i0} they produce the same regression coefficients for the effect of treatment, time and their interaction of treatment and time. If we assume that there are only two repeated measures (a Pre-Post design) it can be shown that:

$$\text{ANCOVA on difference scores} \begin{cases} d_i = \beta_0 + \beta_1 x_{1i} + \beta_2 Y_{i0} + \varepsilon_i \\ Y_{i1} - Y_{i0} = \beta_0 + \beta_1 x_{1i} + \beta_2 Y_{i0} + \varepsilon_i \\ Y_{i1} = \beta_0 + \beta_1 x_{1i} + (1 + \beta_2) Y_{i0} + \varepsilon_i \end{cases}$$

$$\text{ANCOVA on difference scores} \{ Y_{i1} = \beta_0 + \beta_1 x_{1i} + \beta_2 Y_{i0} + \varepsilon_i$$

As can be seen from the above equations, there is no difference between the two alternative models in terms of the treatment effect [328].

9.7 Baseline analysis methods – examples from the literature

The statistical literature is filled with debate regarding the merits and appropriateness of each of these methods. Below we summarise some of the findings from studies comparing alternate baseline analysis strategies

9.7.1 The choice between analysis strategies – ANC vs CSA

The simplest possible longitudinal design comparing two treatment groups consists of a single pre-randomisation baseline measurement and a single post-randomisation response measurement; i.e. $k = 0,1$; where $k = 0$ indexes the pre-randomisation baseline measurement and $k = 1$ indexes the post-randomisation measurement. Such trials are often referred to a Pre-Post designs. A critical feature of Pre-Post designs is that if the baseline measurement is not included as part of the response vector (as is the case with the ANC and CSA analysis strategies) then the resulting model is no longer a repeated measures analysis and simple linear regression techniques can be used to estimate the effect of treatment upon the post-baseline response. Perhaps for this reason, within Pre-Post designs, the treatment effects on mean change from baseline are typically assessed using either an analysis of covariance (ANC) in which the baseline measurement is used as a covariate or using a change from baseline model (CSA). As previously noted in section 9.6.1, the model estimates of the treatment effect from the ANC method would be the same with either the post-baseline value or the change from baseline value as the dependent variable however for simplicity we refer to the former only.

The choice in strategies was first discussed by Frederic M. Lord in 1967 [346] and consequently has become known as Lord's Paradox [346]. In his paper, Lord outlined a hypothetical dataset in which the two groups under comparison differed at baseline and both changed in response equally over time and thus differed by the same amount at the follow-up observation; that is the slope of change was similar between groups. Using, simple regression analysis methods, he noted that the analysis of the change from baseline scores yielded a non-significant result, reflecting the similarity of the change from baseline in both groups, while the ANCOVA indicated

a significant difference between groups at the final assessment. The difference in results may be attributed to the fact that for pre-post designs, the estimate of the treatment effect is equal to the difference at baseline (i.e. the difference in intercepts between groups) plus the difference in change (i.e. the difference in group slopes); or alternatively, the ANCOVA compares only the final difference between groups. Conversely, the estimate of the treatment effect from the change score model is equivalent to the difference in change between baseline and follow-up assessments. The significant result observed in the ANCOVA model may therefore be attributed to the differences observed at baseline while the non-significant results of the change score analysis indicate that once removing the baseline imbalance, both treatment groups changed by a similar amount over the observation period. Despite this result, many statisticians in the health sciences still believe that when baseline measurements are correlated to post-baseline measurements, adjusting for baseline using ANCOVA removes conditional bias in treatment group comparisons due to chance baseline imbalances [347, 348]. Breukelen [349] explored the similarities and differences between the ANC and CSA analysis methods and concluded that the methods differ because ANCOVA assumes absence of a baseline difference between comparator groups. Hedeker and Gibbons [328] note that change score analysis and ANCOVA answer different questions within a Pre-Post Design. Change score analysis tests if the average change is the same between the groups, whereas ANCOVA tests if the post-test average is the same between groups conditional on the baseline mean response being the same between the groups. Despite this conclusion Egbewale et al. [87] compared the bias precision and power of analysis of variance (ANOVA), change-score analysis (CSA) and analysis of covariance (ANCOVA) in response to baseline imbalance. The authors concluded that rather than the ANCOVA being biased in the presence of baseline imbalance, it was the ANOVA and CSA that led to biased estimates, although it should be noted that their definition of bias was difference from the ANCOVA estimate. In 2013 Breukelen again revisited Lord's Paradox in an article aptly titled: "ANCOVA versus change form baseline in nonrandomised studies: The difference" [350]. In this article, the author points out that non-randomised trials do not have the expectation of baseline balance between comparator groups, and as the ANCOVA method is sensitive to baseline imbalance, then it should not be used in nonrandomised studies. Indeed, the perspective that the expectation of baseline balance (i.e. randomisation) is sufficient to justify the use of ANCOVA is

one that has been championed by several prominent statisticians within the health sciences [77, 350, 351].

Given this sensitivity to baseline imbalance it may be hard to understand the widespread popularity of the ANCOVA. Undoubtedly its use relates to two factors: 1) in the absence of baseline imbalance the ANCOVA has the greatest statistical power (efficiency) as compared with any of the 4 analysis methods outlined in the beginning of this section; and 2) change scores are subject to regression to the mean which reduces their efficiency.

Proponents of the ANCOVA have demonstrated through both simulation and analysis of trial data that for pre-post designs, the ANCOVA has the highest statistical power; meaning the ANCOVA is the most statistically efficient method to analyse a continuous outcome with a baseline measurement [244, 352-358]. Adjustment for the baseline covariate reduces the residual variance from $var(Y_{\cdot 1})$ to $var(Y_{\cdot 1}|Y_{\cdot 0})$, the conditional variance of Y after regression on the baseline covariate ($Y_{\cdot 0}$). This can be expressed as $\mathrm{var}(Y_{\cdot 1}|Y_{\cdot 0}) = (1-\rho^2)\mathrm{var}(Y)$, where ρ is the correlation between post- ($Y_{\cdot 1}$) and baseline ($Y_{\cdot 0}$) response measurements [359]. Therefore, the principal effect of covariance adjustment for the baseline response is to increase the precision of treatment contrasts by multiplying the residual variance by the factor $(1-\rho^2)$.

Several authors highlight that when change scores are used to account for differences between groups at baseline, they do not control for baseline imbalances between groups. This is because subjects with low scores at baseline tend to improve more than those with high scores, a phenomenon known as regression to the mean [297]. In general, the variance of a change score can be calculated using the following general equation:

$$\mathrm{var}(\mathrm{D}) = \mathrm{var}(Y_{\cdot 1} - Y_{\cdot 0}) = \mathrm{var}(Y_{\cdot 1})2(1-\rho)$$

Consequently, regression to the mean will increase the variance of the change score relative to the raw response variable when the correlation between pre- and post-measurements is less than 0.5. Therefore, the reduction in the efficiency of the CSA method relative to the ANC is exacerbated by low correlation and ameliorated by high correlation. Vickers [291] undertook simulations to compare differences in statistical power between change scores, percentage

change scores and the analysis of covariance (ANCOVA) and concluded that change scores are an acceptable alternative to ANCOVA if correlation between baseline and post-treatment scores is high.

9.7.2 The choice between analysis strategies – ANC vs cLDA

Liang and Zeger [325] used mixed models to compare ANC and cLDA methods in the analysis of data collected in clinical trials with a pre-post design. The authors demonstrate the similarity of the estimates obtained by regressing the change (post-pre) on the pre-value and on the treatment group and the estimate obtained from a repeated measures analysis in which both the pre- and post-values are considered responses and regressed on a pre-post indicator and pre-post by treatment group interaction. The authors conclude that the cLDA is advantageous as compared with the ANC as the cLDA model can include all randomised subjects with either a baseline value or a postbaseline value and therefore may avoid the bias for estimating within-group mean changes an improve the efficiency for estimating between-group mean differences observed at postbaseline time points [360]. Recently Coffman et al. [361] compared the ANC, cLDA and RMA marginal models in the analysis of pre-post RCT data. The authors support the view that the cLDA may be regarded as the method of choice over ANCOVA and RMA analysis methods.

There are surprisingly few studies that compare baseline analysis strategies when there is more than one follow-up assessment (i.e. when T >1). While Liang and Zeger [325] discuss the applicability of the cLDA approach to multiple post-values and also to discrete data models the article does not contain a systematic comparison of each method's efficiency. Lui et al. [347] compare the ANCOVA model with the cLDA using marginal models in simulations studies for which there are 4 assessment time-points (including baseline). The authors demonstrate that both the ANCOVA and the cLDA models provide unbiased estimate for the treatment difference, although the cLDA model is consistently more powerful than the ANCOVA model. The efficiency loss of the ANCOVA model as compared with the cLDA is partially attributed to treating the baseline values as fixed, which results in an underestimation of the of the mean estimates, where the degree of underestimation depends on both the correlation between baseline and follow-up measurements as well as the variability of the baseline measurement [347]. Lu [360] again revisited the cLDA and longitudinal ANCOVA comparison using simulated data with four repeated measures per subject to explore the efficiency of the two models in the setting of arbitrary missing

data. The author again concluded that if the baseline value is subject to missingness, the cLDA strategy is shown to be more efficient for estimating the treatment differences at postbaseline time points than the longitudinal analysis of covariance.

9.8 Conclusions

Longitudinal studies, employing repeated measurement study design are both common and favoured in the health and medical sciences as well as in pharmaceutical studies. Such studies provide insight into both the ontology of disease and the response profile to new and existing therapies. The past few decades have generated significant advances in the analysis of longitudinal data. Both marginal and mixed models offer trialists tools by which repeated measures data can be analysed to account for intra-subject measurement correlation.

When analysing repeated measures data there appears to be no consensus on how best to incorporate subjects baseline measurements. A significant portion of literature has been devoted to comparing the ANC and CSA baseline analysis strategies within a pre-post design setting however it is unclear if the similarities and differences between the methods are sustained when there are multiple follow-up assessments. While less common, there are a few articles that have explored the cLDA and the ANC in both the pre-post and multi-follow-up settings, however the strength and weaknesses of the RMA and CSA methods in a multi-follow-up setting have not been characterised. To date there has been no systematic review of the analysis strategies across varied correlation strength, structure and number of time assessment time-points. Given that removal of the baseline from the response vector relegates pre-post ANC and CSA analyses to simple linear regression, it not clear how the model estimate interpretations might change with the introduction of time and time by treatment interaction terms that are necessary for longitudinal-ANC and longitudinal-CSA models when $T > 1$. The work in the following chapters seeks to explore some of these gaps in the literature.

In the following two chapters, we explore how both the accuracy (Chapter 10) and efficiency (Chapter 11) are altered depending on how the baseline measurement is incorporated into the analysis of longitudinal data using marginal model analysis methods. Specifically, we will explore the impact of correlation and number of assessment time points upon power and bias.

Chapter 10. Interpreting the regression coefficients – a comparison of statistical methods for the analysis of repeated measures

The greatest value of a picture is when it forces us to notice what we never expected to see.

– John Tukey

10.1 Abstract

Background

The analysis of repeated measures data still represents a challenge for clinical trials collecting repeated measurements on each subject. While general linear models have emerged as the standard for analysing repeated measures data, there are several commonly employed strategies for handling the baseline measurement: 1) retain it as part of the outcome vector (RMA); 2) use the baseline measurement as a covariate in the analysis of the post-baseline measurements (ANC); 3) subtract the baseline measurement from all the post-baseline measurements and then analyse the change scores (CSA). The study objective was to compare the interpretation of parameter estimates from each analysis method as they pertain to known trial parameters.

Methods

Using simulation, we generated five hypothetical trial scenarios involving two treatment arms with an underlying autoregressive structure to the correlation of subject errors. Each trial scenario differed with respect to the baseline imbalance and the pattern of treatment group response. Trial scenarios were then repeated varying the levels of intra-subject correlation and number of subject measurement time points. We then compared each model estimate to the simulation trial parameters.

Results

The regression coefficients from each of the three analysis methods return different trial parameters. The link between trial parameters and model estimates is consistent within the RMA method irrespective of the number of follow-up assessments in the trial design. By contrast, the number of follow-up time points creates alternate interpretation of parameter estimates for both the ANC and CSA analysis strategies. When there is only one follow-up assessment the $time$ and $treatment \times time$ coefficients from the RMA correspond to the $intercept$ and $treatment$ coefficients from the CSA strategy. The $treatment$ coefficient from the ANC strategy captures information about both the intercept and slope difference between treatment arms (i.e. the difference between treatment arms at the final assessment time point). When there is more than one follow-up assessment, all three methods have a similar interpretation of the $treatment \times time$ coefficient - the difference in slope between the two treatment arms.

Conclusions

The interpretation of parameter estimates from all three methods is not equivalent. The degree of concordance partially depends on the number of assessment time points. From an analysis perspective, the distinction between having a single and multiple follow-up assessments is a key feature that should be used to inform the choice of how to handle the baseline measurement.

10.2 Introduction

A longitudinal study refers to a trial or investigation in which patient outcomes are collected at multiple follow-up times. Longitudinal studies might collect data over time from alternate patients (cohort) or from multiple assessments on the same subjects (Figure 10.2.1). If the same subjects are measured repeatedly then the longitudinal study will yield multiple or 'repeated' measurements [74]. Repeated measures experiments are common within medicine and the preferred method of design for clinical trials evaluating new interventions. In repeated measures clinical trials, following screening, subjects are randomised to receive either a treatment or a control. They then are given a baseline evaluation, receive their assigned treatment, and are then followed for a period of time during which they will receive one or more assessments pertaining to the outcome variable relating to the indication under investigation. The distinction between having a single or multiple follow-up assessments is a key feature of our results and discussion and therefore we refer to experiments in which there is a single follow-up as a 'pre-post' [303, 362], and experiments in which there are two or more follow-up assessments as 'multi-follow-up'. In both cases, the primary efficacy measure is often the pattern or trajectory of change from baseline to the last pre-specified visit assessment in which the outcome/response variable is measured.

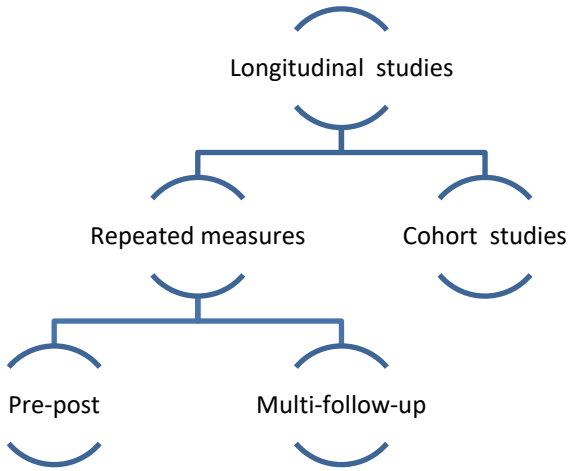


Figure 10.2.1 Different subcategories of longitudinal studies

The analysis of repeated measures data still represents a challenge for clinical trials collecting repeated measurements on each subject. While general linear models (both mixed and marginal) have emerged as the standard for analysing repeated measures data, there is still a lack of

consensus about how to handle the baseline value. The statistical properties of baseline adjustment methods are complex and often poorly understood [363]. Much of the discussion surrounding the choice of analysis is rendered inaccessible to all but a handful of subject matter experts as the problem is usually discussed within the context of Lord's Paradox [346]. A literature review [364] of 50 clinical trials from top medical journals (e.g. British Medical Journal, Journal of the American Medical Association, The Lancet, and New England Journal of Medicine) revealed the use of several alternate baseline adjustment methods. It is our contention that the lack of consistency, or at very least a justification of the chosen method stems from the fact that there is little guidance for the overwhelming majority of researchers performing longitudinal analyses. Furthermore, the lack of consistency in the literature on the analysis of baseline measures further obfuscates the relative merits of each method.

In this paper we consider three ways of handling the baseline value (adapted from [326]):

1) Retain the baseline measurement as part of the outcome vector (RMA);
2) Use the baseline measurement as a covariate in the analysis of the post-baseline measurements (ANC)
3) Subtract the baseline measurement from all of the remaining post-baseline measurements and then analyse the change scores (CSA).

It is our contention that many researchers analysing repeated measures have limited understanding of the implications for alternate methods of baseline adjustment. Moreover, there is a lack of appreciation for how baseline adjustment methods interact with intra-patient measurement correlation and frequency of patient assessment.

Therefore, our goal is to describe, discuss, and demonstrate how to interpret the regression coefficients from each analysis strategy. We simulated hypothetical trial scenarios by levels of: baseline imbalance; treatment effect; within-subject correlation and number of follow-up measurements. We analyse each trial scenario using each of the three baseline adjustment methods above and then compare the regression coefficients against the trial design parameters.

10.2.1 Notation

We continue with the notation outlined in the previous Chapter 9, i.e. we let Y_{ik} represent a response observed at time k for observation $k = 0, \dots, T$ on subject $i = 1, \dots, N$. For scalar representation x_{1i} is used to denote the treatment group assignment for the i^{th} subject (x_{1i}=0 for control and x_{1i}=1 for active). x_{2k} is used to denote the specific time point of assessment, hence

$x_{2k} = 0$ for the baseline measurement and $x_{2k} = T$ for the time at final assessment for $m = T + 1$ follow-up time points in n patients.

As much of this work has to do with the comparison of treatment group means we continue the notation from section 7.2.1 and represent the mean response for treatment group j at time k as μ_{jk}. To aid in the presentation of our results we define the intercept (group mean at baseline) of the control and active arms as: $\mu_{00} = I_0$ for intercept of the control group and $\mu_{01} = I_1$ intercept of the active group. The difference between groups at baseline is $D_0 = I_1 - I_0$. Similarly, the mean difference between treatment arms at final assessment is $D_T = \mu_{1T,} - \mu_{0T}$. In contrast to the previous chapter, we now consider baseline imbalance, the effect of which is that $D_T \neq \Delta$, instead we reserve effect size (Δ) to denote the difference in slopes between treatment groups ($\Delta = S_1 - S_0$), where the slope is calculated as $S_j = (\mu_{jT} - \mu_{j0})/T$. Our perspective is that increased patient assessments (m) result in more frequent patient assessment rather than prolonged patient assessment and hence when exploring trials with different numbers of assessment time points the slope is kept constant between trial scenarios by reducing the time between assessment points (see Figure 10.2.2).

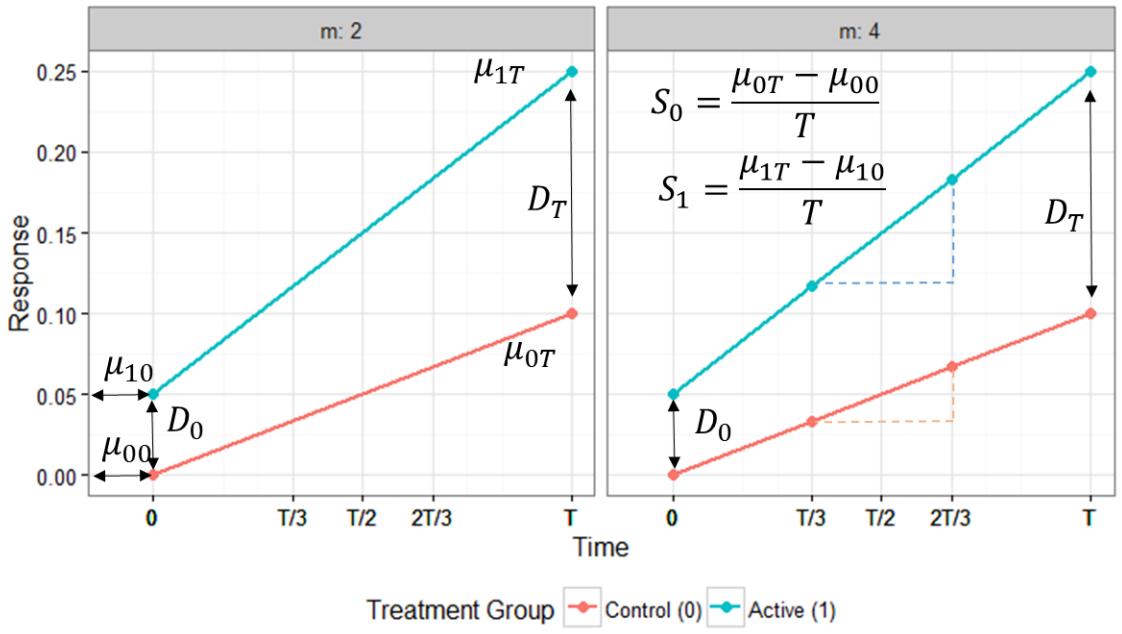


Figure 10.2.2 Hypothetical trial scenarios.
m:2 has only has one follow-up measurement, m:4 has three follow-up measurements. Each panel has the same initial difference (D_0), final difference - effect size (Δ) for the treatment groups. The treatment specific slopes (S_j) are calculated from $S_j = (\mu_{jT} - \mu_{j0})/T$. Consequently, by reducing the interval between assessments, more frequent follow-up, the treatment specific slopes remain constant across panels.

10.2.2 Baseline adjustment analysis options

We consider three analytical approaches before/after data that are commonly used: 1) repeated measures analysis including the baseline value for each patient in the response vector (RMA), 2) analysis using the post treatment administration response as the outcome and using the baseline measurement as a covariate or predictor (ANC), and 3) analysis of difference scores (CSA). Below we detail the algebraic form of each strategy.

10.2.3 Analysis strategies as regression models

10.2.3.1 Repeated measures analysis (RMA)

In the repeated measures analysis, the baseline value of the response variable in a series of repeated measures may be modelled simply as one of the repeated outcome measures:

$$y_{ik} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik} \quad (10.2.1)$$

Where y_{ik} is the response score for person i $(i = 1, \dots, n)$ at time point k $(k = 0, \dots, T)$, (0=pre and T = final assessment), and ε_{ik} is the patient by time error matrix.

10.2.3.2 Analysis of covariance (ANC)

In the repeated measures analysis of covariance, the baseline value of the response variable is modelled as a covariate. The principle of this analysis is to adjust the scores of the dependent variable considering the influence of the baseline covariate.

This model differs from the RMA is two ways: 1) each patient now has a vector of responses that does not include baseline $(k = 1, \dots, T)$ and 2) in addition to the intercept (β_0), treatment (β_1), time (β_2) and treatment by time interaction (β_3), the model contains a parameter for the baseline measurement $(\beta_4 y_{ij0})$: When there is more than one follow-up assessment (i.e.$m > 2$) the ANCOVA can be written in regression form as follows:

$$y_{ik|k>0} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \beta_4 y_{i0} + \varepsilon_{ik} \quad (10.2.2)$$

Where y_{ik} is the response score for person i $(i = 1, \dots, n)$ at time point t $(t = 1, \dots, T)$. Note that the vector of time points now ranges from $1, \dots T$ instead of $0, .. T$ as a consequence of the baseline

becoming a covariate. For a pre-post design, the inclusion of baseline as covariate results in the removal of time (x_{2k}) and $treatment \times time\ (x_{1i}x_{2k})$ terms from the model:

$$y_{i1} = \beta_0 + \beta_1 x_{1i} + \beta_4 y_{i0} + \varepsilon_{i1} \quad (10.2.3)$$

Note that even though there only 3 coefficients in the model we still denote the coefficient associated with the baseline response measurement as β_4.

10.2.3.3 Change score analysis (CSA)

In the analysis of change score the baseline measurement for each patient is subtracted from each post-baseline response measurement. Similar to the ANC, the principle of this test is to adjust for potential baseline imbalance, however, unlike the ANC approach this model does not include the baseline covariate on the right-hand side of the equation:

$$y_{ik|k>0} - y_{i0} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + e_{ik} \quad (10.2.4)$$

Similar to the ANC, for the CSA method, the vector of response time points now ranges from $1, \ldots T$; and hence for a pre-post design, the model does not include a $time\ (x_{2k})\ or\ treatment \times time\ (x_{1i}x_{2k})$ term in the model:

$$y_{i1} - y_{i0} = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad (10.2.5)$$

10.2.4 Covariance

Correlation between measurements on the same subject result in a covariance structure imposed upon the errors during the estimation of model coefficients (this is covered in more detail in the next chapter). A general covariance structure is denoted as:

$$Cov(Y_{ik}, Y_{il}) = Cov(\varepsilon_{ik}, \varepsilon_{il}) = \begin{bmatrix} \sigma^2{}_k & \sigma_{kl} \\ \sigma_{kl} & \sigma^2{}_l \end{bmatrix} \quad (10.2.6)$$

Where σ_{kl} is the covariance between measurements at times k and l on the same subject, and $\sigma_{kk} = \sigma_k^2$ is the variance at time k. Assuming a fixed common variance between time points k and l this can be rewritten as:

$$\begin{bmatrix} {\sigma^2}_k & \sigma_{kl} \\ \sigma_{kl} & {\sigma^2}_l \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho_{kl} \\ \rho_{kl} & 1 \end{bmatrix} \quad (10.2.7)$$

10.2.4.1 Autoregressive (AR(1))

Autoregressive covariance structure specifies that covariances between intra-patient response observations decrease with increasing distance between observations [343, 365]. This is to say that observations on the same patient are increasingly independent when farther apart.

As our simulation work fixed the correlation between the first and last time point then the correlation between successive time points can be calculated from the exponential function:

$$\rho_{kl} = \rho^{\left(\frac{|k-l|}{m-1}\right)} \quad (10.2.8)$$

Where k and l are time points of subject observation, m is the total number of subject assessment time points (including baseline) and ρ is the desired correlation level between the baseline and final assessment time points. Table 10.2.1 shows an example AR(1) correlation matrix for $\rho = 0.25$ and $m = 4$.

Table 10.2.1 AR(1) intra-patient response correlation matrix between time points for ρ=0.25 and m=4

		Time point l			
Time point k		**0**	**1**	**2**	**3**
	0	1	0.63	0.40	0.25
	1	0.63	1	0.63	0.40
	2	0.40	0.63	1	0.63
	3	0.25	0.40	0.63	1

10.3 Methods

10.3.1 Data simulation

To compare the performance of the three strategies for handling baseline scores in a marginal model repeated measures analysis, we carried out a simulation study. The primary aim of this study was to assess the coefficients of the RMA, ANC and CSA methods under alternate trial conditions. Within the simulations we created alternate trial conditions by manipulating: the number of assessment time points, baseline imbalance, difference effect size and pre-post correlation strength. Model coefficients from the analysis were then recorded for each approach, under each condition.

Each dataset is based on the settings described in 10.3.2 using the following assumptions: 1) each dataset is based on two treatment groups (active and control coded as 1 and 0 respectively), 2) all data are completely observed for all patients, 3) all measurement time points are equally spaced, 4) the variance at each time point was equal between treatment groups and constant across time, 5) each treatment arm contained 10,000 subjects to minimise error from the model estimates. We chose this balanced situation so that we can illustrate the model coefficients from each analysis strategy relative to data design without complications introduced by missing, unbalanced data or sample size limitations. All data generation were performed using SAS v.9.4. We used the SAS MIXED procedure (SAS Institute Inc. 2008) within SAS STAT [154] for all marginal model analyses. Appendix E contains portions of the SAS code used in the analyses.

10.3.1.1 Data simulation process

The data simulation consisted of the following steps:

1) A correlation matrix (C) was constructed for each trial scenario (see Table 10.3.1).
2) Using Cholesky decomposition a matrix was identified such that $U^T U = C$.
3) Random sampling from a $N(0,1)$ distribution was used to generate an $n\ x\ m$ matrix of uncorrelated errors (E_u).
4) We generated correlated errors (E_c) from uncorrelated (E_u) by multiplication with the U matrix $(E_c. = E_u U)$.
5) The time point specific responses for each patient were calculated using the treatment specific intercepts and slopes from the design Table 10.3.1.

10.3.2 Design parameters

Table 10.3.1 lists the values of trial parameters used in the simulated datasets. These trial parameters ($\boldsymbol{I_0, I_1, S_0, S_1}$) , describe the degree to which the two simulated treatment arms differ at baseline ($\boldsymbol{D_0}$), and also difference between treatment arm response over time Δ. The value of the intercept and slope varied between 0 and 0.25 for both treatment arms. As the variance was set to one for both treatment groups and all time points, then the values of the intercepts and effect sizes (Δ) are equivalent to a Cohen-D effect size. The value of the intercept and slope varied between 0 and 0.25 for both treatment arms. As the variance was set to one for both treatment groups and all time points, then the values of the intercepts and effect sizes (Δ) are equivalent to a Cohen-D effect size.

Table 10.3.1 Simulated parameters

Design Number	$\boldsymbol{I_0}$	$\boldsymbol{I_1}$	$\boldsymbol{S_0}$	$\boldsymbol{S_1}$	$\boldsymbol{D_0 = I_1 - I_0}$	$\Delta = \boldsymbol{S_1 - S_0}$
1	0	0.25	0.25	0	0.25	-0.25
2	0	0.25	0	0	0.25	0
3	0	0.25	0	0.25	0.25	0.25
4	0	0	0.25	0	0	-0.25
5	0	0	0	0	0	0
6	0	0	0	0.25	0	0.25
7	0.25	0	0.25	0	-0.25	-0.25
8	0.25	0	0	0	-0.25	0
9	0.25	0	0	0.25	-0.25	0.25

Figure 10.3.1 provides a visual representation of the mean treatment response for each simulated data set. Each parameter set (design number) was repeated for five different intra-patient correlation levels (0,0.25,0.50, 0.75 and 0.99), and three different number of assessment time points (2, 4 and 6) using an AR(1) correlation structure.

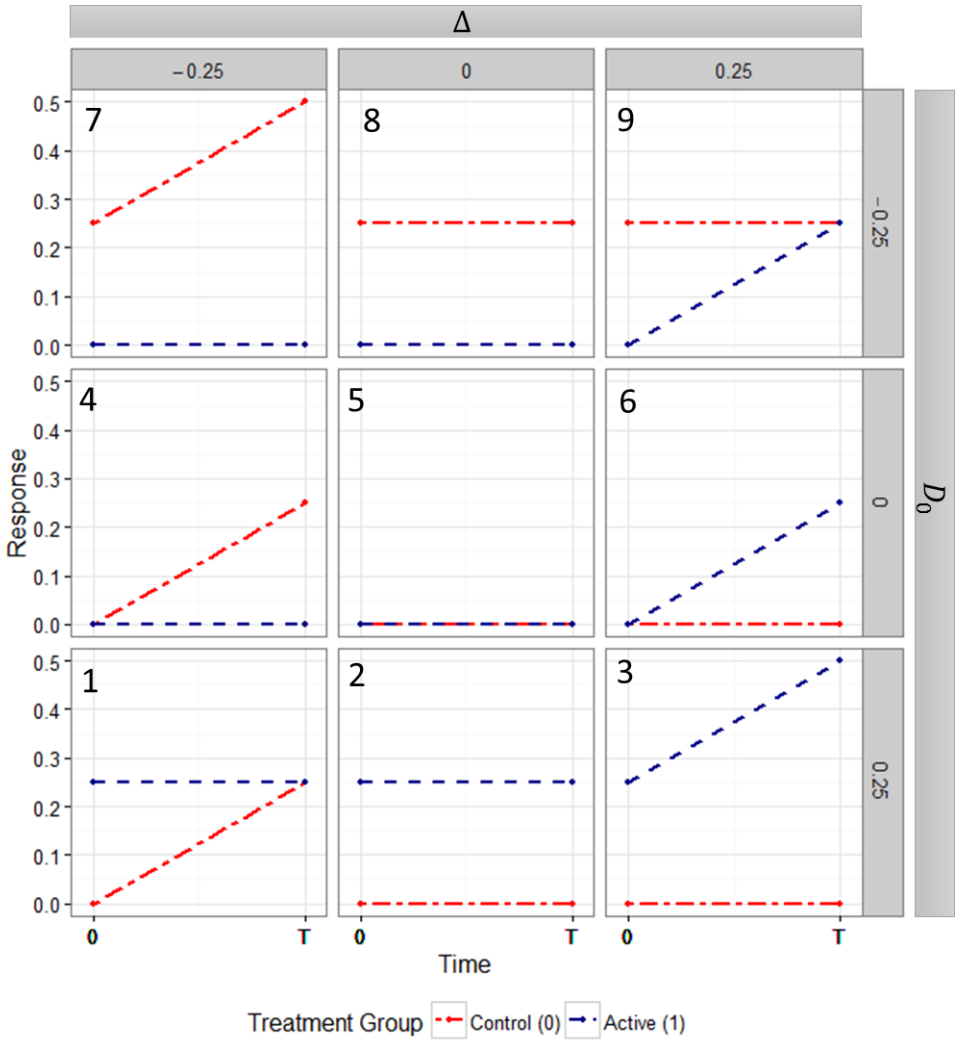


Figure 10.3.1 Shows the mean treatment group response over time for each simulated dataset. Parameter set is mirrored across the diagonal reversing treatment group assignment to each trial parameter.

10.4 Results

Figure 10.4.1-Figure 10.4.5 show the estimates returned for example simulated datasets. The examples have been chosen to represent treatment arm differences from the following four categories: 1) null, 2) baseline imbalance, 3) effect difference and 4) combination of effect difference and baseline imbalance (see Table 10.4.1).

Table 10.4.1 Summary table highlighting the design parameters used in the example figures

Associated Figure number	Design number	Design Parameters	Difference between treatment arms	
			At baseline	Effect size
10.4.1	5	$I_0 = 0,\ I_1 = 0,\ S_0 = 0,\ S_1 = 0$	No	No
10.4.2	6	$I_0 = 0,\ I_1 = 0,\ S_0 = 0,\ S_1 = 0.25$	No	Yes
10.4.3	1	$I_0 = 0,\ I_1 = 0.25,\ S_0 = 0.25,\ S_1 = 0$	Yes	Yes
10.4.4	2	$I_0 = 0,\ I_1 = 0.25,\ S_0 = 0,\ S_1 = 0$	Yes	No
10.4.5	3	$I_0 = 0,\ I_1 = 0.25, S_0 = 0,\ S_1 = 0.25$	Yes	Yes

Table 10.4.2 summarises the relationship between the beta coefficient and the design parameters. Two columns of statistics are provided for each analysis model. Firstly, an interpretation of the results for when there are more than two patient assessment time points (i.e. 4 or 6). The second grouping of statistics provides the estimate interpretation when there are only 2 patient assessment time points. Below we detail each coefficient for each analysis method.

10.4.1 RMA

The RMA analysis method has a similar interpretation of the regression coefficients irrespective of the number of time points. The intercept coefficient (β_0) captures the control arm intercept (I_0); and the $treatment$ coefficient (β_1) is the difference in intercept between the two treatment arms ($I_1 - I_0$). Similarly, the time coefficient (β_2) is the slope of the control arm (S_0) while the coefficient for the $treatment \times time$ interaction (β_3) estimates the difference in slopes between treatment arms ($S_1 - S_0$).

10.4.2 CSA

When there are more than two assessment time points, the CSA analysis method returns a zero estimate for the intercept and treatment coefficients (β_0 and $\beta_1$respectively). The time coefficient

captures the slope of the control arm (S_0) and the $treatment \times time$ interaction coefficient (β_3) estimates the difference in slopes between treatment arms ($S_1 - S_0$).

When there are only two assessment time points (baseline and one follow-up) then the $intercept$ coefficient (β_0) captures the slope of the control arm (S_0) while the $treatment$ interaction coefficient (β_1) estimates the difference in treatment arm slopes ($S_1 - S_0$).

10.4.3 ANC

Table 10.4.3 and Figure 10.4.6 both present the average correlation between baseline and all subsequent time points for the null model. The expected value was calculated by summing the correlation between baseline and each subsequent measurement and then dividing by the number of post baseline measurements (see Appendix F). The expected and observed values are identical when there are only 2 time points but the observed value is fractionally higher than the expected value for both 4 and 6 time points.

The value of the baseline coefficient plays an important role in the estimation of the regression coefficients for both intercept and treatment. When there are more than two time points, at zero correlation, the $intercept$ coefficient is equal to the intercept for the control arm (I_1). As the correlation increases this value decreases proportionally to the increase in the $baseline$ coefficient multiplied by the intercept of the control group (I_1) such that at perfect correlation between baseline and subsequent measurements the value of $intercept$ would be zero. Similarly, at zero correlation and more than two time points, the $treatment$ coefficient is equal to the difference in intercepts between the treatment arms ($I_1 - I_0$) with increases in correlation resulting in a decrease in the coefficient magnitude proportional to the increase in $baseline$ multiplied by the difference in intercepts (($I_1 - I_0) \times baseline$).

When there is only a baseline and one follow-up assessment (2 time points), at zero correlation, the $intercept$ coefficient is equal to the sum of the intercept and slope for the control group ($S_0 + I_0$). Increases in correlation result in a decrease in the coefficient value proportional to the baseline coefficient multiplied by the intercept for the control arm ($I_0 \times baseline$). At zero correlation, the treatment coefficient is equal to the sum of the slope and intercept differences between treatment arms (($S_1 - S_0) + (I_1 - I_0)$). An increase in correlation results in a decrease

in the treatment coefficient equal to the $baseline$ coefficient multiplied by the difference in treatment arm intercepts ($baseline \times (I_1 - I_0)$).

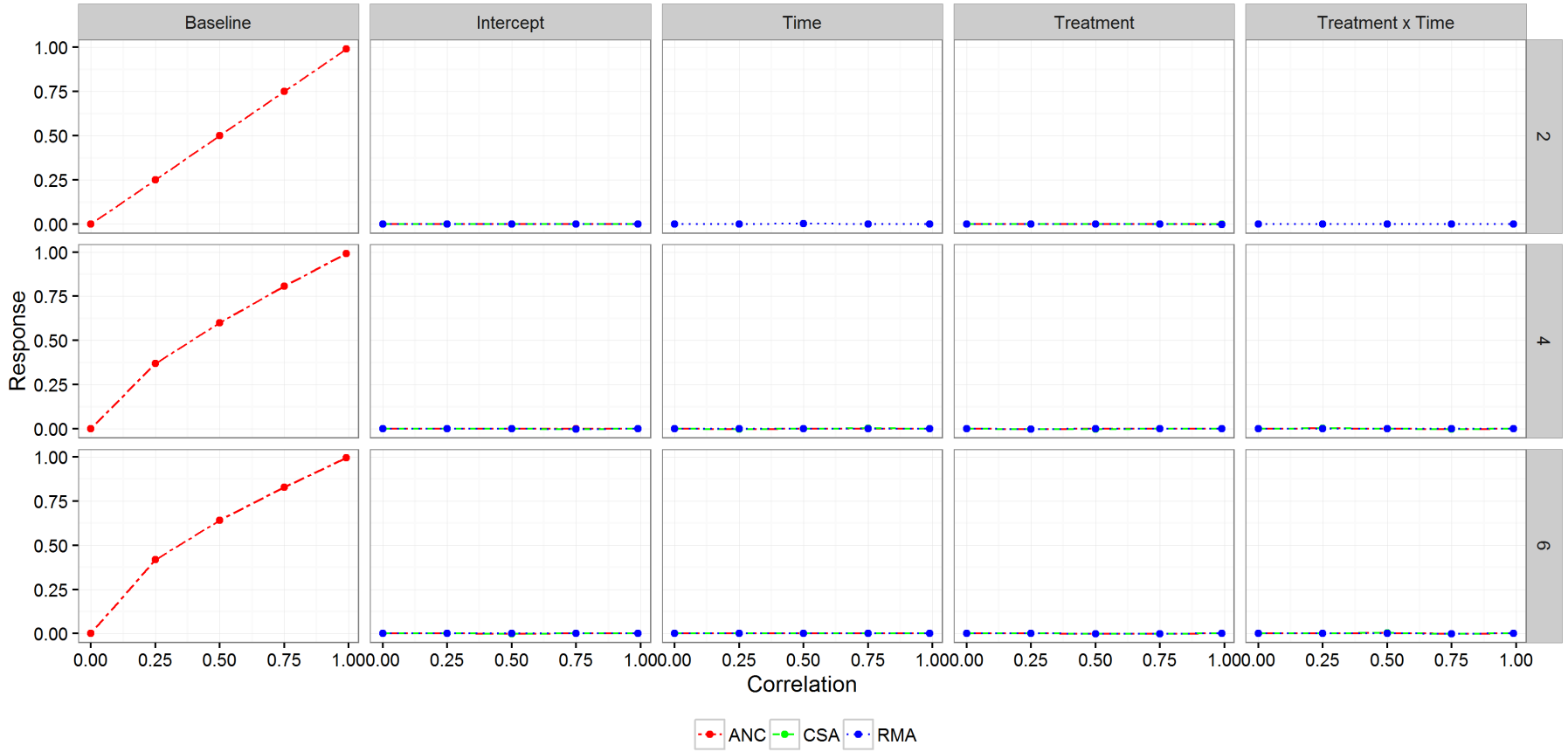


Figure 10.4.1 Observed model coefficients against correlation from design number 5.

Datasets with a no baseline imbalance and null slope difference: $I_0 = 0,\ I_1 = 0,\ S_0 = 0,\ S_1 = 0$. The baseline coefficient from the ANC strategy increases with correlation. The pattern of increase varies against the number of time points. When there are two time points, the magnitude of the baseline coefficient increases linearly with correlation. When there are more than two time points the increase in the baseline coefficient is less than the increase in correlation with a 'diminishing returns' pattern.

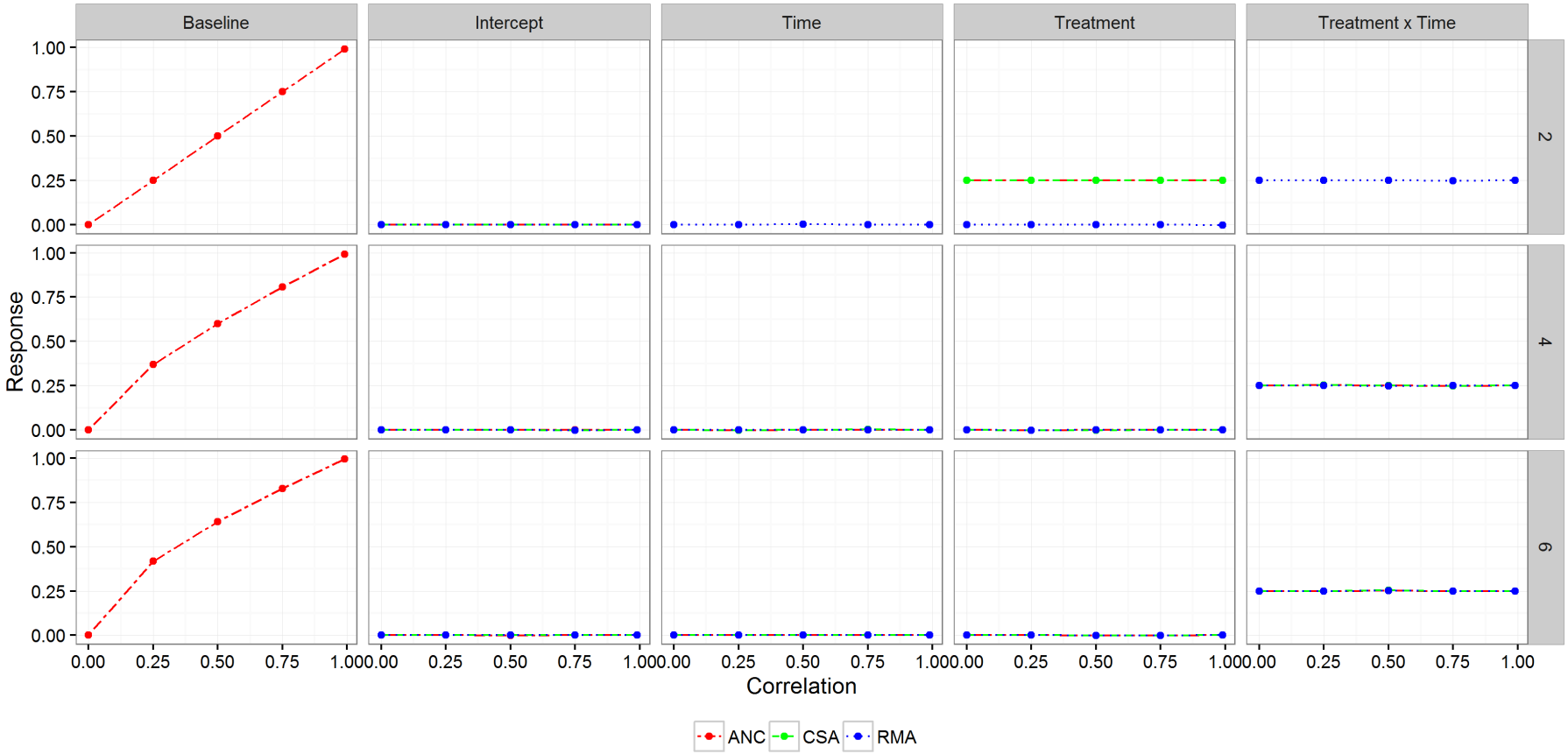


Figure 10.4.2 Observed model coefficients against correlation from design number 6.

Datasets with no baseline imbalance and a positive slope difference between treatment arms: $I_0 = 0,\ I_1 = 0,\ S_0 = 0,\ S_1 = 0.25$. When there are 2 time points the difference between treatment slopes is captured by the $treatment$ coefficient for both the ANC and CSA methods and by the $treatment \times time$ coefficient for the RMA method. When there are more than 2 time points, the slope difference is captured by the $treatment \times time$ interaction for all three baseline strategies (RMA, ANC, CSA).

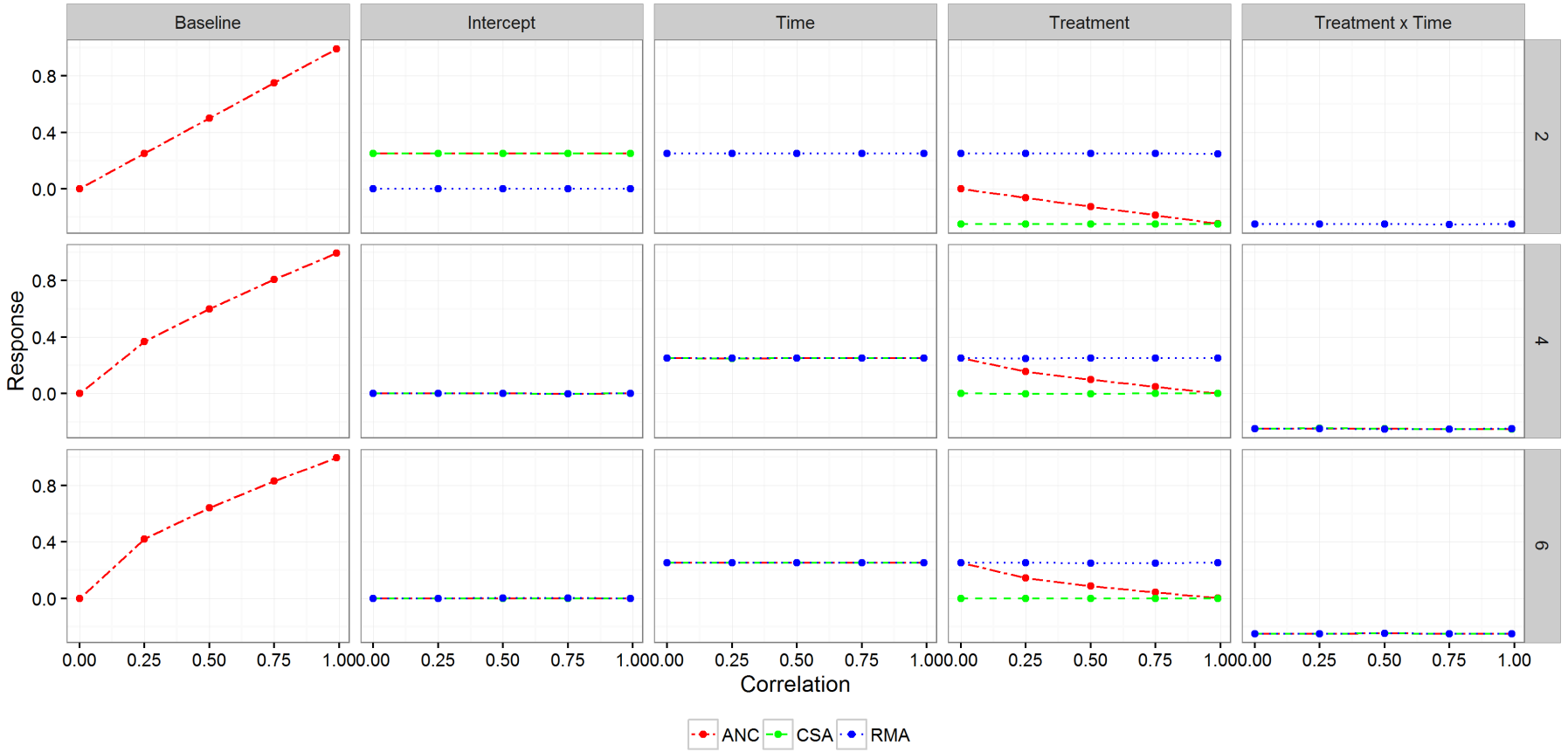


Figure 10.4.3 Observed model coefficients against correlation from design number 1.
Datasets with a positive intercept difference (the active group starts with higher values than the control group) and a negative slope difference (the control group has a slope while the active group does not change over time): $I_0 = 0,\ I_1 = 0.25,\ S_0 = 0.25,\ S_1 = 0.$ The treatment coefficient from the ANC method decreases with increasing correlation.

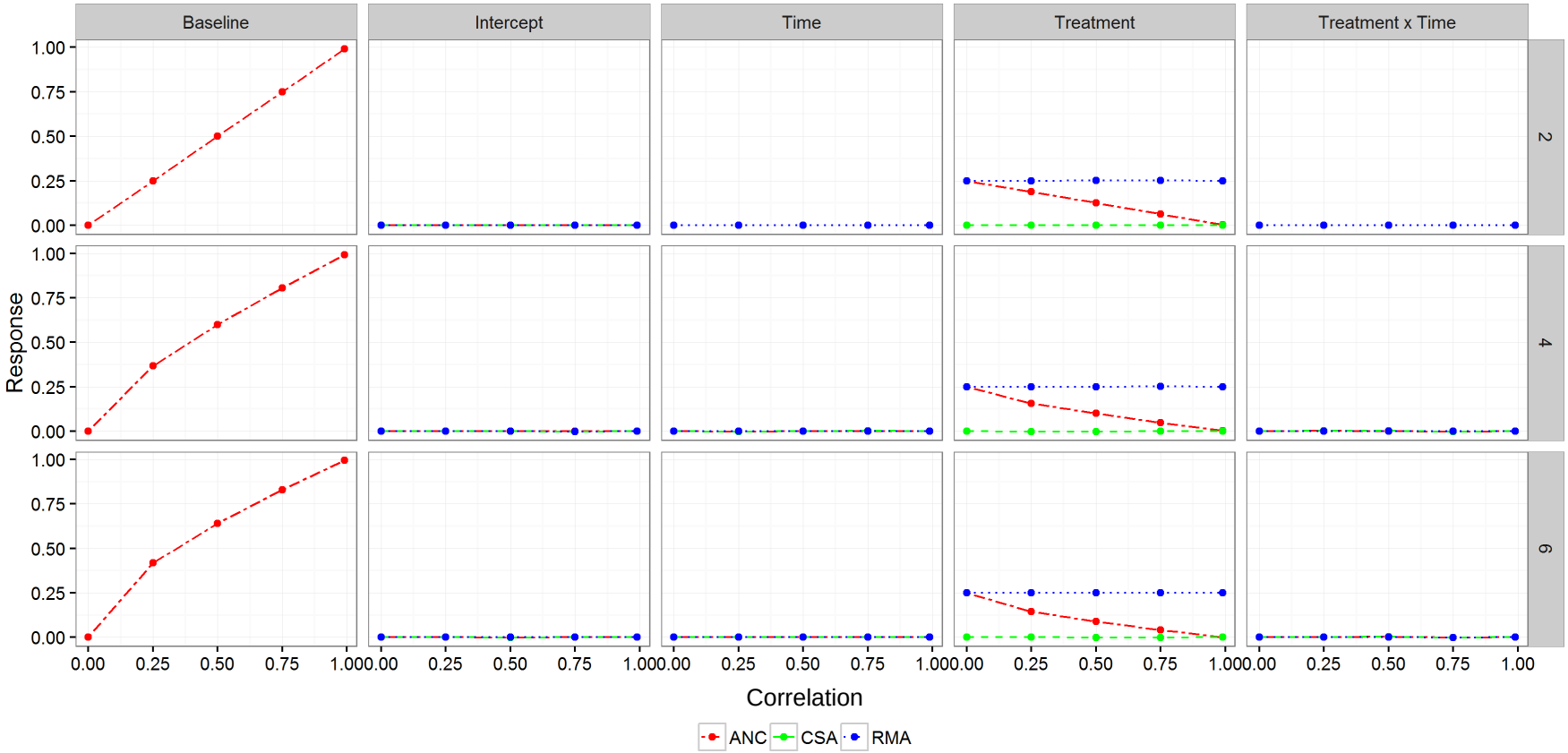


Figure 10.4.4 Observed model coefficients against correlation from design number 2.
Datasets with a positive intercept difference (the active group starts with higher values) and a null treatment effect (both active and control groups have a zero slope): $I_0 = 0,\ I_1 = 0.25,\ S_0 = 0,\ S_1 = 0.$ The treatment coefficient from the ANC method decreases with increasing correlation. The specific value of the ANC treatment coefficient is between the values observed in the RMA ($\rho = 0$) and the CSA ($\rho = 1$) analyses.

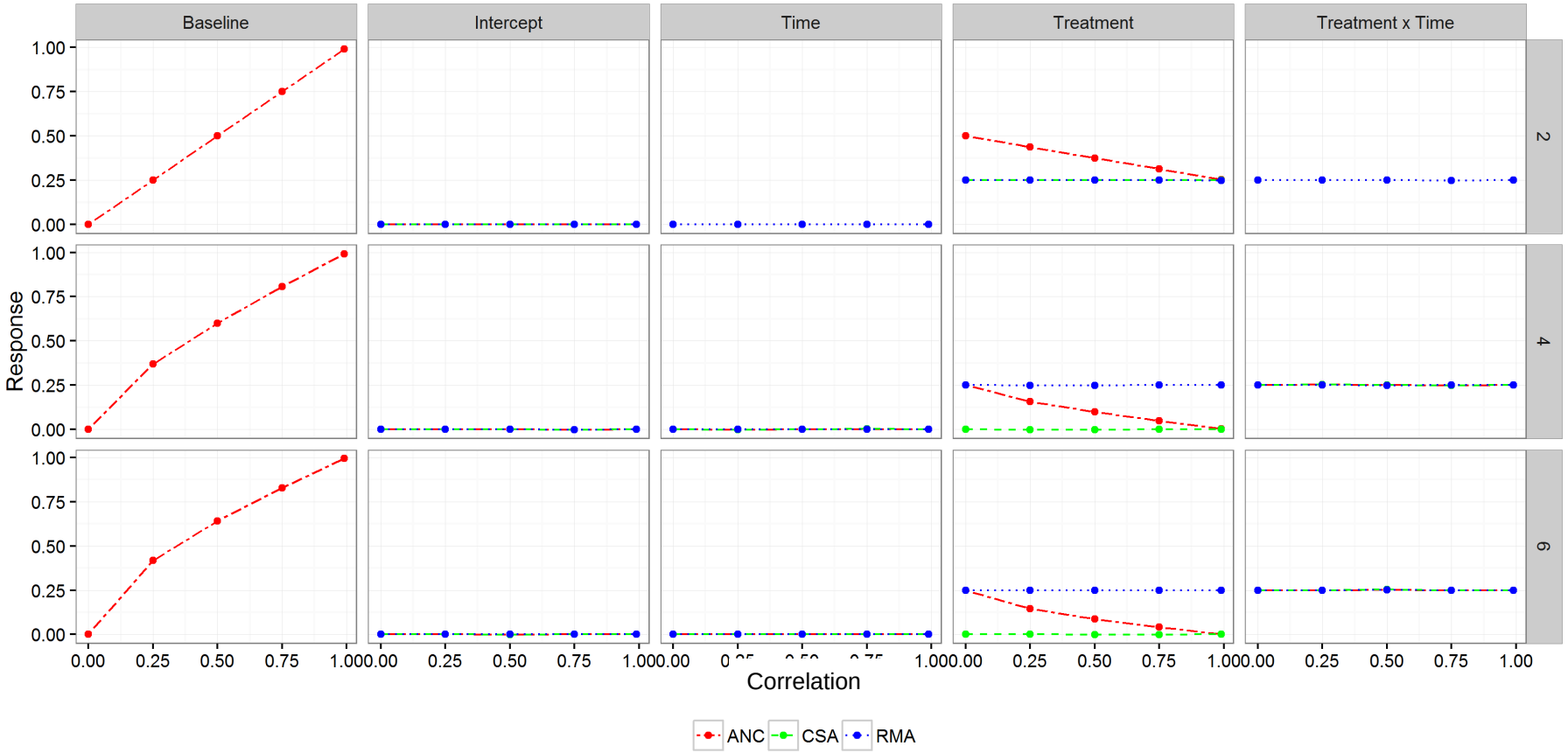


Figure 10.4.5 Observed model coefficients against correlation from design number 3.
Datasets with a positive intercept difference (the active group starts with higher values) and a positive slope difference (active has a positive slope and control group has a zero slope): $I_0 = 0,\ I_1 = 0.25,\ S_0 = 0,\ S_1 = 0.25$. The treatment coefficient from the ANC method decreases with increasing correlation. The specific value of the ANC treatment coefficient is between the values observed in the RMA ($\rho = 0$) and the CSA ($\rho = 1$) analyses.

Table 10.4.2 Regression coefficients interpretation

Parameter	Coefficient	Pre - Post			Multi-follow-up		
		RMA	**ANC**	**CSA**	**RMA**	**ANC**	**CSA**
Intercept	β_0	I_0	$S_0 + I_0(1-\beta_4)$	S_0	I_0	$I_0(1-\beta_4)$	0
Treatment	β_1	$I_1 - I_0$	$(S_1 - S_0) + (I_1 - I_0)(1-\beta_4)$	$S_1 - S_0$	$I_1 - I_0$	$(I_1 - I_0)(1-\beta_4)$	0
Time	β_2	S_0	N/A	N/A	S_0	S_0	S_0
Treatment x Time	β_3	$S_1 - S_0$	N/A	N/A	$S_1 - S_0$	$S_1 - S_0$	$S_1 - S_0$
Baseline $(y_{\cdot\cdot 0})$	β_4	N/A	Average intra-patient correlation†	N/A	N/A	Average intra-patient correlation†	N/A

† see Table 10.4.3

Table 10.4.3 The expected† and observed average intra-patient correlation between baseline and all subsequent measurements for the null model

Correlation	$m = 2$			$m = 4$			$m = 6$		
	Expected	**Observed**	**Error**	**Expected**	**Observed**	**Error**	**Expected**	**Observed**	**Error**
0	0	1.01e-16	7.072e-3	0	1.16e-17	4.083e-3	0	3.85e-18	3.162e-3
0.25	0.25	0.25	6.847e-3	0.426	0.432	5.077e-3	0.469	0.486	4.706e-3
0.5	0.50	0.50	6.124e-3	0.641	0.644	4.536e-3	0.673	0.679	4.192e-3
0.75	0.75	0.75	4.677e-3	0.828	0.829	3.42e-3	0.844	0.846	3.154e-3

The expected average correlation is calculated as $\frac{1}{T}\sum_{k=0}^{T}\rho_{0k}$.

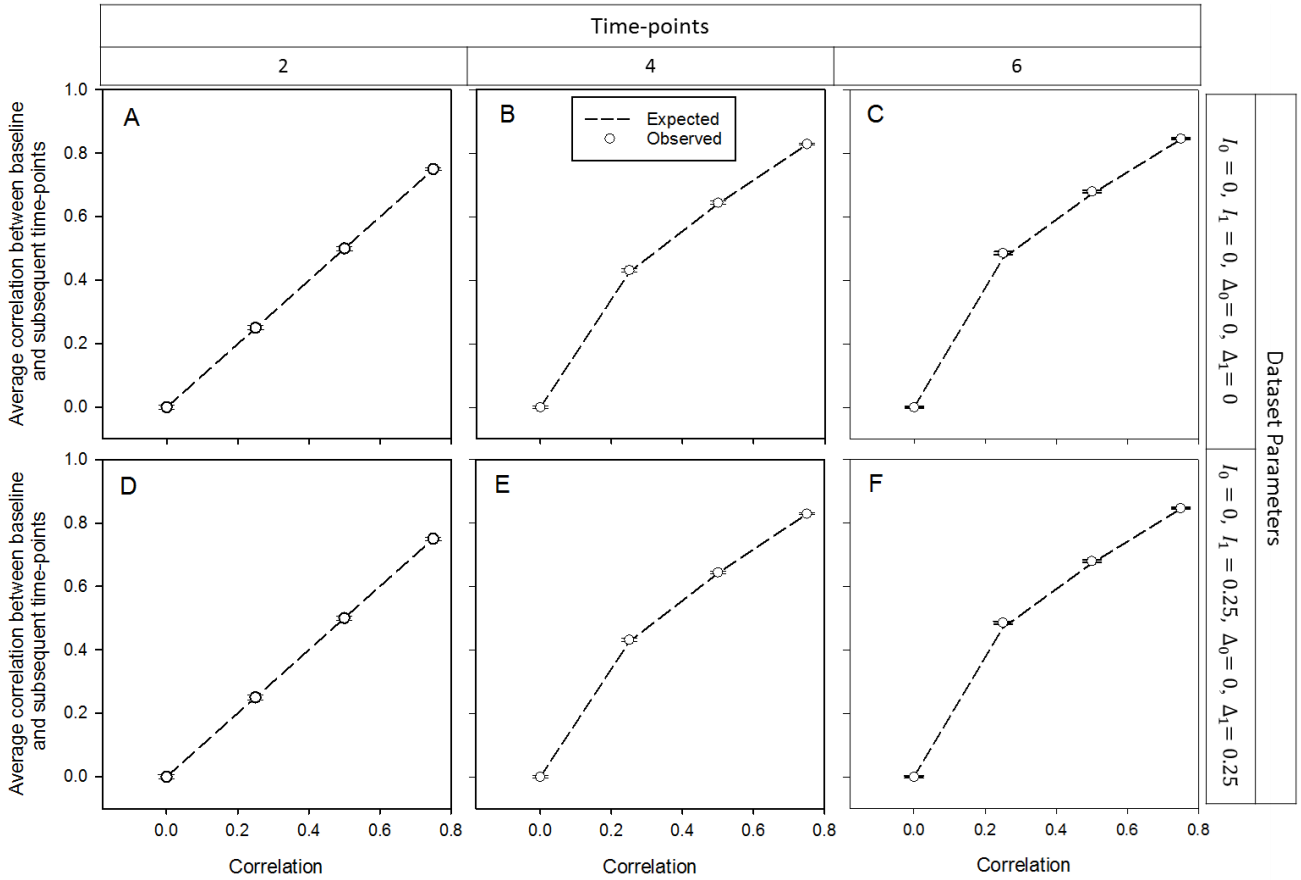


Figure 10.4.6 The expected and observed average intra-patient correlation between baseline and all subsequent measurements.
Panels A, B, and C are datasets that have the same 'null' design parameters ($I_0 = 0,\ I_1 = 0,\ \Delta_0 = 0,\ \Delta_1 = 0$) but differ in the number time points of assessment. Similarly, panels D, E and F all have the same 'effect' design parameters ($I_0 = 0,\ I_1 = 0.25,\ \Delta_0 = 0,\ \Delta_1 = 0.25$) but differ in the number of time points. The clear circles in each graph show the observed β estimates relating to the baseline parameter at each level of correlation. The dashed line shows the expected average correlation between the baseline values and all subsequent time points (see Appendix D).

10.5 Discussion

This simulation study has examined the effect of baseline imbalance and number of patient assessments on the interpretation of model coefficients as they pertain to known trial parameters. We revealed that the regression coefficients of $treatment$ from each of the three analysis methods relate to alternate trial parameters. Even within the same analysis method, both CSA and ANC coefficients describe alternate trial parameters depending on if there is one or more than one follow-up assessment. When there is only one follow-up assessment (pre-post) the $treatment$ coefficient of the CSA measures the difference in slopes between while the $treatment$ coefficient of the ANC measures the: slope difference + (baseline imbalance – correlation between pre-post measurements × baseline imbalance). The discrepancy between the interpretation of the treatment coefficient between change scores and inclusion of the baseline score as a covariate has been discussed for a long time in statistical literature [77, 346, 350, 366, 367], and is most frequently presented as Lord's Paradox [346]. Essentially the argument recognises that the ANC method reports as the measure of treatment effect the group difference at final assessment by definition the combination of baseline imbalance and slope difference between treatment arms). It has been argued [77] that this does not negate the validity of the treatment effect measured by the ANC method as the observation of baseline imbalance is not the same as 'true' baseline imbalance. While baseline imbalance may reflect a prognostic factor unaccounted for in the randomisation, it may also occur by chance alone. If we accept the latter is the case, then as a normally distributed random process sometimes the baseline imbalance will 'increase' the power of the ANCOVA and sometimes it will 'decrease' the power; it follows that across all the many trials employing this analysis method the technique is 'unbiased'. Consequently, to employ the ANC in the analysis of repeated measures data requires only the *expectation* of no baseline imbalance rather than the *observation* of no baseline imbalance. This paradigm has resulted in the widely accepted recommendation that the ANCOVA is an appropriate test for randomised trials [297, 368], where randomisation would achieve the expectation of baseline balance, and an inappropriate test procedure for observation trials where there can be no expectation of baseline balance between comparator groups owing to the lack of randomisation [368, 369]. In reality, the 'expectation' of no imbalance is very hard to achieve

owing to limits on the number of prognostic factors incorporated into the randomisation (see Introduction), and it may be hard to fully justify the expectation of baseline balance when analysing and presenting the results of an individual trial. Proponents of ANC point out that the change scores used in the CSA analysis method are subject to 'regression to the mean', a phenomenon that refers to the tendency for subjects who have below average pre-scores tend to have higher post-scores, and for those above average at pre-test tend to have lower post-scores. Regression to the mean alters the variance of change scores proportional to the intra-subject correlation strength, if the correlation is low using the change score will add variation and the treatment coefficient is less likely to show a significant result. Conversely, if the correlation is high the change score will have reduced variance and the treatment coefficient is more likely to be significant [297].

While we are not the first authors to discuss the discrepancy between ANC and CSA coefficient interpretation we hope that this work has helped elucidate the impact that correlation between pre- and post- measurements has upon the treatment estimate in the ANC model. Moreover, we believe that describing the model estimates relative to trial parameters may aid researchers in the interpretation of model coefficients.

Where there is more than one follow-up measurement (multi-follow-up trial design), the interpretation of the model coefficients for both the ANC and CSA is altered. Within the multi-follow-up setting, the $Treatment \times Time$ interaction coefficient measures the slope difference between treatment arms across all three analysis methods (RMA, ANC and CSA). Similarly, in the multi-follow-up setting the $Treatment$ coefficient captures the difference in treatment arms for both the RMA and ANC methods, although it should be noted that as the ANC treatment estimate is multiplied by $(1 - \beta_4)$, then it will be smaller than the treatment estimate from the RMA when the average correlation between baseline and follow-up assessments is greater than zero. For the CSA method, the trial parameters described by the $intercept$ and $treatment$ terms in a pre-post design is instead described by the $time$ and $treatment \times time$ interaction terms in a multi-follow-up design.

In contrast to the CSA and ANC methods, the RMA has consistent interpretation of regression coefficients irrespective of the number of follow-up measurements (pre-post and multi-follow-up)

and accurately characterises any baseline imbalance (treatment coefficient) as well as the difference in slopes ($treatment \times time$ coefficient) between comparator groups. This consistency allows for the direct comparison of model coefficients between pre-post and multi-follow-up trials. Moreover, if we accept that a regression estimates should describe the data then only the RMA method allows for recreation of the mean treatment responses at baseline and over time from the model coefficients, both the ANC and CSA methods do not allow for data visualisation.

A statistical interaction occurs when the effect of one independent variable on the dependent variable changes depending on the level of another independent variable. A $treatment \times time$ interaction is equivalent to asking whether the expectation of change in response over time changes depending on the treatment received. If the change over time (slope) of effect for the control group is different from the effect of time on active, then there is an interaction. As standard statistical practice is to ignore the 'significance' of a main effect in the presence of an interaction between them it would appear that all three methods offer similar results when there is more than one follow-up assessment. This assertion of course ignores power, and it is important to note both the ANC and CSA analysis methods exclude the baseline measurement from the response vector of each patient. In the next chapter, we explore the effect that this has upon the power of studies to meaningfully detect significant trial effects. In recent published literature, the use of ANCOVA as the analysis method of choice for determining the treatment effect while controlling for any baseline imbalance has been advocated by several authors [297, 349, 353, 354, 370, 371]. Our results do not support this view and we are not sure that perceived definition of 'control' for baseline imbalance matches the expectation of many researchers who employ this analysis method to their data. The addition of baseline as a covariate produces a coefficient that approximates the average correlation between time points. While this would conceivably reduce the residual variation (a topic explored further in the next chapter), we do not observe evidence that inclusion of the baseline as a covariate produces more accurate estimates of patient change over time between groups, which we believe is the intended effect for the majority of researchers.

To summarise, in this section we have reviewed three strategies to handle the baseline measurement in the analysis of longitudinal data. The RMA strategy retains the baseline value as part of the outcome. This method has consistent interpretation of regression coefficients

irrespective of the number of follow-up measurements. The other two baseline strategies (ANC and CSA) do not use the baseline value as part of the outcome. Both of these methods have alternate interpretation of regression coefficients depending on the number of follow-up measurements. Moreover, the ANC method does not separate information regarding the baseline and slope difference between treatments arms when there is one follow-up measurement. We conclude that in both randomised and observational trials, the RMA methods should be the analysis strategy of choice.

Chapter 11. Baseline as a covariate or dependent variable in standard error of the model estimates

All models are wrong, but some are useful.

– George E. P. Box

11.1 Abstract

Background

Analysis of variance (RMA), change-score analysis (CSA) and analysis of covariance (ANC) are three commonly employed methods for the analysis of repeated measures data. Numerous studies have reported the conditions in which all three methods yield unbiased coefficient estimates however few studies have quantified the differences between methods in the calculation of the standard error of the coefficient estimates, particularly in a multi-follow-up design. As the standard error has direct implications for precision of estimates derived from these methods of analysis it defines their relative statistical power.

Methods

We consider the design and covariance matrices from each baseline analysis method in pre-post and multi-follow-up scenarios. We explore how the alternate parametrisation of the baseline measurement impacts the variance of the model estimates. Simulation results are presented for support of theoretical calculations for example scenarios

Results

For pre-post studies, the estimated treatment effect from the ANC is more precise than the estimated treatment effect from the RMA and CSA methods. Consequently, the ANC is more powerful than the RMA and CSA methods. The difference in power between methods is ameliorated by high intra-patient correlation between pre- and post- measurements. For multi-follow-up studies, the precision of the treatment effect is greatest in the RMA method. The RMA has markedly improved power compared with the ANC and CSA when the correlation structure of the data is compound symmetry and marginally improved power when the correlation structure is autoregressive.

Conclusions

Researchers need to carefully consider the number of assessment time points and covariance structure when selecting an analysis method for repeated measures data.

11.2 Introduction

Much of the research in medical science is based upon longitudinal designs that involve repeated measurements of a variable of interest from each participating patient. Such experiments usually collect a baseline value of the response variable and one or more response variable assessments after a treatment intervention. In repeated measures clinical trials, a crucial analytic decision is whether to adjust for the baseline measurement of the response variable.

In this section, we continue to review procedures for the inclusion of baseline in the analysis of repeated measures data; these are:

1) Include the baseline in the outcome vector (RMA);
2) Remove the baseline measurement from the outcome vector and include it as a covariate (ANC);
3) Subtract the baseline measurement from each subsequent measurement for each patient and then analyse the change scores (CSA).

The decision of which method to employ is driven by two statistical concerns 1) Bias and 2) Power. If treatment groups are not similar at baseline then it may introduce bias in the form of an estimate of the treatment effect that deviates from its true value [372]. The ANC model is one of the most commonly used statistical methods for the analysis of change for pre-post designs and there is a wealth of statistical literature establishing the equality of ANC and CSA estimates [77, 87, 349, 373]. In Chapter 9 we established the validity of this claim on the assumption that baseline values are unbiased between treatment groups. Articles advocating the use of ANC for the analysis of pre-post trials frequently point out that the change scores used in CSA analysis are subject to regression to the mean, and hence the ANC is the more powerful method as the residual variance is decreased through inclusion of the baseline as a covariate [77, 296, 297, 374].

Generally speaking, the power of the test represents the probability of detecting differences between the groups being compared when such differences exist. When testing regression coefficients, it is convenient to consider power as a consequence of: 1) the effect size, and 2) standard error of estimation.

The relationship between these two aspects is expressed clearly by the Wald test. For any given coefficient β the Wald statistic is given by the formula [374]:

$$\frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}} = t \approx z_{1-\alpha} + z_{1-\beta} \quad (11.2.1)$$

Where $\hat{\beta}$ is the estimate of the model coefficient, which is approximately normally distributed with standard error $\left(\sqrt{var(\hat{\beta})} = SE(\beta)\right)$. The standard deviation of the estimate of a regression coefficient measures how precisely the model estimates the coefficient's unknown value. Smaller values of the standard error indicate a more precise estimate. The Wald test is used to test the null hypothesis that a regression coefficient β is 0:

$$H_0: \beta = 0 \quad (11.2.2)$$

Thus, for a fixed type I error ($z_{1-\alpha}$), the power will be high if the effect size β is large and/or the standard error is small.

In Chapter 9 we established the conditions that yield equality in the accuracy of the model coefficients however even when each baseline analysis method yields similar coefficient estimates this does not mean that all three baseline analysis strategies will have equivalent power. The focus of this section is on comparing the precision of alternate baseline analysis strategies by comparing how each strategy calculates variance of the estimated model parameters (i.e. $SE(\hat{\beta})$). To accomplish this, we have several objectives:

The first is to review the analysis of longitudinal data. The second is to illustrate the decomposition of longitudinal models to their component matrices. The third is to use these matrices to calculate the theoretical $SE(\beta)$ over each baseline analysis strategy. As our previous work established that the interpretation of the regression coefficients differs between pre-post and multi-follow-up designs, we separate these within our work here. Much of the focus of this chapter is on the covariance matrix and the similarities and differences between baseline analysis strategies that yield differences in statistical power.

The chapter is organised as follows.

1) Preliminary material about longitudinal models is recapped in Section 11.3.
2) Section 11.4 reviews $SE(\beta)$ calculations for pre-post experiments.
3) Section 11.5 reviews $SE(\beta)$ calculations for multi-follow-up experiments.
4) Section 11.6 presents the results of a simulation study comparing the accuracy of the formulae reported in Sections 11.4 & 11.5.
5) Concluding remarks appear in Section 11.7.

11.3 Linear models for longitudinal data

11.3.1.1 Notation

We continue the notation outlined in section 7.2.1. and 9.3.1.

Let Y_{ik} denote the value of the response measured at time k on subject i; $i = 1, \dots, N$, and $k = 0, \dots, T$. Again, as time is coded with baseline as 0, there are T post baseline assessment measurements and $T + 1 = m$ total number of assessments time points. For scalar representation x_{1i} is used to denote the treatment group assignment for the i^{th} subject (x_{1i}=0 for control and x_{1i}=1 for active). Much of this work has to do with covariance between assessment time points and to aid in presentation we refer to separate time points using k and l. A general covariance structure is denoted as $cov(Y_{ik}, Y_{il}) = \sigma_{kl}$, where σ_{kl} is the covariance between measures at time k and l, and $\sigma_{kk} = \sigma_k^2$ is the variance at time k. The mean of Y_{ik} is represented by $E(Y_{ik})$. Lastly, we use i and h to refer to separate subjects.

11.3.2 Simple linear regression (SLR)

As covered previously in section 9.3.1, in matrix notation, the regression equation for the *i*th subject takes the form:

$$Y_i = X_i\beta + \varepsilon_i \qquad (11.3.1)$$

Where X_i is an $n_i \times p$ matrix of predictor variables, $\beta_c, c = 0,1, \dots, p$ are unknown regression coefficients (where β_0 is the intercept). In a simple (or multivariable) linear regression, it is assumed that $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and $Cov(\varepsilon_i, \varepsilon_h) = 0$.

With n independent subjects, we can organise the regression equations for each sample unit into a single model:

$$Y = X\beta + \varepsilon \tag{11.3.2}$$

Where Y is an $n \times 1$ vector of subject responses, β is an $n \times (p+1)$ matrix of predictor variables, and ε is an $n \times 1$ vector of random errors for Y.

For this model, it is assumed that $E(Y) = X\beta$ and $v(Y) = \sigma^2 I$:

$$Y \sim N(X\beta, \sigma^2 I) \tag{11.3.3}$$

One common approach to estimate the β vector is to choose a value of β that minimises the sum of the squared residuals, this is commonly known as ordinary least squares (OLS).

$\hat{\beta}$ is used to denote the least squares estimate of β. The formula for deriving this estimated vector is:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{11.3.4}$$

Where $E(\hat{\beta}) = \beta$, and:

$$Var(\beta) = \sigma^2(X'X)^{-1} \tag{11.3.5}$$

The variance-covariance matrix of the OLS estimator, $\hat{\beta}$ is given by [375]:

$$var(\beta) = (X^T X)^{-1} = \begin{pmatrix} var(\beta_0) & cov(\beta_0, \beta_1) & \cdots & cov(\beta_0, \beta_p) \\ cov(\beta_1, \beta_0) & var(\beta_1) & \cdots & cov(\beta_1, \beta_p) \\ \vdots & cov(\beta_2, \beta_1) & \ddots & cov(\beta_2, \beta_p) \\ cov(\beta_p, \beta_0) & cov(\beta_3, \beta_1) & \cdots & var(\beta_p) \end{pmatrix} \tag{11.3.6}$$

Where the standard errors of each model coefficient are given by the square root of the elements along the main diagonal of (11.3.6):

$$SE(\beta) = \sqrt{diag(X^T X)^{-1}} \tag{11.3.7}$$

11.3.3 Multivariate linear regression with a marginal model

A consequence of repeated measures designs is that measurements made on the same individual are often correlated. As discussed in section 9.3, marginal models are an extension of the linear model that account for repeated response measurements from the same study participant [74]. The difference between the marginal model and a simple linear regression model is that the residuals ($\boldsymbol{\varepsilon}$) are no longer assumed to be independent [74, 375, 376], instead the variance is assumed to be $Var(\varepsilon_i) = R_i$. Similar to simple regression, the residuals are still expected to sum to zero, $E(\varepsilon_i) = 0$. These two assumptions are given by:

$$\varepsilon \sim N(0, \mathrm{R}) \tag{11.3.8}$$

The matrix R is an $(n \cdot m) \times (n \cdot m)$ block diagonal matrix with repeating subunits blocks R_i showing the covariance of errors from the repeated measurements across time points for each patient.

$$\mathrm{R} = \begin{bmatrix} R_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_N \end{bmatrix}_{(N \cdot m) \times (N \cdot m)} \tag{11.3.9}$$

Where each patient subunit R_i shows the population averaged covariance between time points at which measurements are collected. It follows that the dimensions of the R_i matrix are defined by the number of patient measurement time points:

$$R_i = \begin{bmatrix} \sigma_{00} & \sigma_{01} & \cdots & \sigma_{0T} \\ \sigma_{10} & \sigma_{11} & \cdots & \sigma_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T0} & \sigma_{T1} & \cdots & \sigma_{TT} \end{bmatrix}_{(m) \times (m)} \tag{11.3.10}$$

$$= Cov(\varepsilon_{ik,}\varepsilon_{il}) = cov(y_{ik,}y_{il})$$

As correlation is simply the variance adjusted covariance, the R_i can be further decomposed into common correlation matrix of repeated measures estimated across all subjects.

$$\mathrm{R_i} = \sigma^2 \begin{bmatrix} 1 & p_{01} & \cdots & p_{0T} \\ p_{10} & 1 & \cdots & p_{1T} \\ \vdots & & \ddots & \\ p_{T0} & & & 1 \end{bmatrix} \tag{11.3.11}$$

As the correlation matrix is symmetric (see section 7.3.2.2) it follows that the R matrix is also symmetric by design. We will return to discuss covariance structures in section 11.5.2 but for now it is only important to note that, as the covariance is a function of correlation, the specific elements of R_i will vary based upon the correlation of errors between time points and that marginal models directly estimate the correlations among each individual's residuals with the assumption that the residuals across different individuals are independent of each other [74, 377].

Owing to the correlation of measurements from the same patient, the estimation procedures from equations (11.3.4) and (11.3.5) are inadequate to estimate the model coefficients and their variance as they fail to account for the intra-patient correlation of response measurements. Within a marginal model analysis framework, information regarding intra-patient measurement correlation strength and structure is incorporated into parameter estimation in a repeated measures analysis through the inclusion of the covariance matrix R such that:

$$\hat{\beta} = (X'R^{-1}X)^{-1}X'\mathrm{R}^{-1}Y \tag{11.3.12}$$

And similarly, the variance of the estimates is calculated as:

$$Var(\beta) = (X'\mathrm{R}^{-1}X)^{-1} \tag{11.3.13}$$

As with simple linear regression, the standard errors of each model coefficient are given by the square root of the elements along the main diagonal of (11.3.6):

$$SE(\beta) = \sqrt{diag(X^T\mathrm{R}^{-1}X)^{-1}} \tag{11.3.14}$$

11.4 Pre-post designs

In this section, we compare analytic methods for experimental data when there is a single baseline and a single follow-up measurement – pre-post design. We focus on the situation where the comparison is between two groups: x_{1i}= 0 denotes membership to the control group; and x_{1i}= 1 denotes that the patient is a member of the active or intervention group. Furthermore, we assume

that there is no missing data and thus for each subject i we have a baseline measurement denoted as y_{i0} and a follow-up measurement denoted as y_{i1}.

Table 11.4.1 presents the scalar forms of each baseline analysis option for pre-post data.

Table 11.4.1 Analysis options for baseline using regression methods for a two group comparison of pre-post data

Baseline analysis method	Scalar model
RMA	$Y_{ik} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2k} + \beta_3 x_{1i} x_{2k} + \varepsilon_{ik}$
ANC	$Y_{i1} = \beta_0 + \beta_1 x_{1i} + \beta_4 Y_{i0} + \varepsilon_{i1}$
CSA	$Y_{i1} - Y_{i0} = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

As we highlighted in our previous chapter, each of these regression models provides parameters with different interpretations. Specifically, for the ANC and CSA analyses the coefficient β_1 represents the difference in the mean response at follow-up comparing $\boldsymbol{x_{1i}}$= 1 to $\boldsymbol{x_{1i}}$= 0. For the RMA analysis, the coefficient β_3 captures the slope difference between treatment arms and is therefore the coefficient of interest as it captures treatment arm differences with respect to 'change' from baseline.

11.4.1 Scalar forms → matrix forms

Using the scalar models from Table 11.4.1 we can form the design and response matrices corresponding to each analysis strategy (RMA, ANC and CSA). Table 11.4.2 shows the design and response matrices for the RMA, ANC and CSA analyses in a pre-post design. As both the ANC and CSA methods remove time from the model they have half the number of rows as the RMA in their design matrices and response vectors. Additionally, the inclusion of $time$ $(\boldsymbol{x_{2k}})$ and $time\ X\ treatment$ $(\boldsymbol{x_{1i}x_{2k}})$ as model parameters results in the RMA having a larger design matrix than either then ANC or CSA methods.

Table 11.4.2 Pre-post design and response matrices

Analysis Strategy	Design Matrix – $\boldsymbol{X}$	Response Vector - $\boldsymbol{Y}$
RMA	$\begin{bmatrix} 1 & x_{11} & x_{20} & x_{11}x_{20} \\ 1 & x_{11} & x_{21} & x_{11}x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{20} & x_{1N}x_{20} \\ 1 & x_{1N} & x_{21} & x_{1N}x_{21} \end{bmatrix}_{(N\cdot 2)\times 4}$	$\begin{bmatrix} y_{10} \\ y_{11} \\ \vdots \\ y_{N0} \\ y_{N1} \end{bmatrix}_{(N\cdot 2)\times 1}$
ANC	$\begin{bmatrix} 1 & x_{11} & y_{10} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & y_{\mathrm{N0}} \end{bmatrix}_{N\times 3}$	$\begin{bmatrix} y_{11} \\ \vdots \\ y_{N1} \end{bmatrix}_{N\times 1}$
CSA	$\begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1N} \end{bmatrix}_{N\times 2}$	$\begin{bmatrix} y_{11} - y_{10} \\ \vdots \\ y_{\mathrm{N1}} - y_{\mathrm{N0}} \end{bmatrix} = \begin{bmatrix} \mathrm{D}_1 \\ \vdots \\ \mathrm{D}_N \end{bmatrix}_{N\times 1}$

11.4.2 Correlation ($\boldsymbol{\rho}$), covariance ($\boldsymbol{R}$) and inverse-covariance ($\boldsymbol{V}$) matrices

As stated previously, the appropriate analysis of repeated measures data requires that the within-subject correlation be accounted for. A marginal model accomplishes this through the introduction of a variance-covariance matrix (hereafter referred to as just the covariance matrix) when estimating the variance of each of the model coefficients ($\beta's$). The covariance captures the variance adjusted correlation between of the response vector time points through estimation of the covariance between errors of assessment time points.

Within the context of a pre-post design, the covariance matrix estimates the covariance of errors between the pre- and post-response vectors (Y_0 and Y_1). Assuming a common variance to both time points, we can construct the variance-covariance matrix using the following steps.

Let R be a block diagonal matrix of $(n \cdot 2) \times (n \cdot 2)$ showing the sample correlation between pre- and post-response vectors:

$$\mathrm{R}_{(N\cdot 2)\times(\mathrm{N}\cdot 2)} = \sigma^2 \begin{bmatrix} \rho_{00(1)} & \rho_{01(1)} & 0 & 0 & \cdots & 0 & 0 \\ \rho_{10(1)} & \rho_{11(1)} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \rho_{00(2)} & \rho_{01(2)} & \cdots & 0 & 0 \\ 0 & 0 & \rho_{10(2)} & \rho_{11(2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho_{00(N)} & \rho_{01(N)} \\ 0 & 0 & 0 & 0 & 0 & \rho_{10(N)} & \rho_{11(N)} \end{bmatrix} \quad (11.4.1)$$

Where each repeating subunit: $R_i = \sigma^2 \begin{pmatrix} \rho_{00(i)} & \rho_{01(i)} \\ \rho_{10(i)} & \rho_{11(i)} \end{pmatrix}$, therefore:

$$R^{-1} = \begin{bmatrix} (R_1)^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (R_N)^{-1} \end{bmatrix} \quad (11.4.2)$$

If we denote the inverse of each $(R_i)^{-1}$ as V_i then:

$$R^{-1} = V = \begin{bmatrix} V_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_N \end{bmatrix} \quad (11.4.3)$$

Where each V_i submatrix is composed of 4 elements denoting the inverse covariance between pre- and post- measurements:

$$V_i = \begin{bmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{bmatrix}$$

In which the number of rows ($k, k = 0,1$); and the number of columns ($l, l = 0,1$) is equal to m.

If we assume that the variance of error at each time point is 1 then the covariance matrix is simply the inverse correlation matrix. To illustrate the effect of correlation, Table 11.4.3 presents correlation matrices and their inverse for a range of correlation values (0 – 0.99) between pre- and post - measurements. The first two columns present correlation and corresponding inverse covariance assuming an error variance of 1 for both pre- and post- response vectors. From the inverse covariance matrix, we define three properties, which will be useful in later work: 1) the sum of all the elements in the covariance matrix (w), 2) the variance of the post- response vector (u_{11}), and 3) the ratio of the post- variance and the sum of all covariance elements $\frac{u_{11}}{w}$.

Table 11.4.3 Correlation and covariance

Correlation/ Covariance $(\boldsymbol{R_i})$	**Inverse Covariance** $(\boldsymbol{V_i})$	$\sum_{k,l=0}^{1} \boldsymbol{u_{kl}} = \boldsymbol{w}$	$\boldsymbol{u_{11}}$	$\frac{\boldsymbol{u_{11}}}{\boldsymbol{w}}$
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	2	1	$\frac{1}{2}$
$\begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}$	$\begin{bmatrix} 1.067 & -0.267 \\ -0.267 & 1.067 \end{bmatrix}$	1.6	1.067	$\frac{2}{3}$
$\begin{pmatrix} 1 & 0.50 \\ 0.50 & 1 \end{pmatrix}$	$\begin{bmatrix} 1.333 & -0.667 \\ -0.667 & 1.333 \end{bmatrix}$	1.3333	1.333	1
$\begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}$	$\begin{bmatrix} 2.286 & -1.714 \\ -1.714 & 2.286 \end{bmatrix}$	1.142	2.286	2
$\begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$	$\begin{bmatrix} 50.25 & -49.75 \\ -49.75 & 50.25 \end{bmatrix}$	1.005	50.25	50

11.4.3 Inverse-covariance $(\boldsymbol{V})$ matrices – ANC, CSA and RMA

For the RMA it is obvious that:

$$V_i = \begin{bmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{bmatrix}$$

However, as both ANC and CSA have the post time point in the outcome vector then:

$$V_i = (\sigma^2 \boldsymbol{I})^{-1}$$

However, even if the variance of each response vector (Y_0 and Y_1) are equal to 1, the variance of the error in both the CSA and ANC will ≠ 1. Below I will discuss the reasons for each.

11.4.3.1 CSA – regression to the mean

Regression towards the mean refers to the statistical phenomenon whereby individuals with high pre-scores will tend to move down on the post-test, while individuals with low pre-test scores will tend to move up [296, 378, 379]. As a consequence, Y_0 will be negatively correlated with Y_1 and the reduction in variance of the change score will be proportional the strength of the correlation between pre and post measurements [297]:

$$\text{var}(Y_1 - Y_0) = \sigma_{1-0}^2 = \sigma_0^2 + \sigma_1^2 - 2\rho_{01}\sigma_0\sigma_1 \qquad (11.4.4)$$

In repeated measures analysis, estimation of the standard error of the coefficients uses the inverse of the residual, which yields:

$$V_i = \frac{1}{\sigma_0^2 + \sigma_1^2 - 2\rho_{01}\sigma_0\sigma_1} \tag{11.4.5}$$

To show how this quantity compares with the information presented in Table 11.4.3, let us assume that $\sigma_0^2 = \sigma_1^2 = 1$, equation (11.4.5) is reduced to:

$$V_i = \frac{1}{2(1 - \rho_{01})} \tag{11.4.6}$$

Table 11.4.4 shows the change in the variance of a pre-post test score across a range of correlation values. It can be seen from the table that (see Appendix G):

$$V_i = \frac{1}{\sigma_0^2 + \sigma_1^2 - 2\rho_{01}\sigma_0\sigma_1} = \frac{u_{11}}{w} \tag{11.4.7}$$

Table 11.4.4 V_i matrix of the CSA method for a pre-post design

Correlation Matrix	$\boldsymbol{\rho_{01}}$	$\frac{\mathbf{1}}{\mathbf{2(1-\rho_{01})}}$	$\frac{\boldsymbol{u_{11}}}{\boldsymbol{w}}$
$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	0	$\frac{1}{2}$	$\frac{1}{2}$
$\begin{pmatrix}1 & 0.25\\0.25 & 1\end{pmatrix}$	0.25	$\frac{2}{3}$	$\frac{2}{3}$
$\begin{pmatrix}1 & 0.50\\0.50 & 1\end{pmatrix}$	0.50	1	1
$\begin{pmatrix}1 & 0.75\\0.75 & 1\end{pmatrix}$	0.75	2	2
$\begin{pmatrix}1 & 0.99\\0.99 & 1\end{pmatrix}$	0.99	50	50

11.4.3.2 ANC – variance reduced

To explain the variance calculation let us redefine the covariance matrix of the RMA with the following form:

$$\mathrm{R}_i = \begin{bmatrix}\sigma_a^2 + \sigma_e^2 & \sigma_a^2\\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2\end{bmatrix} = \begin{bmatrix}\sigma^2\rho_{00} & \sigma^2\rho_{01}\\ \sigma^2\rho_{10} & \sigma^2\rho_{11}\end{bmatrix} = \begin{bmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{10} & \sigma_1^2\end{bmatrix} \tag{11.4.8}$$

Equation (11.4.8) reveals that the total variance at pre- (σ_0^2) or post- (σ_1^2) can be partitioned into residual variance (σ_e^2) and correlation variance (σ_a^2) ; where σ_a^2 is the within subject covariance created by the correlation, and residual or error variance (σ_e^2) is the variance ignoring the intra-subject correlation [380].

For the ANC (and CSA), there is only 1-time point, so the covariance matrix is simply:

$$\mathrm{R}_i = [\sigma_1^2] \tag{11.4.9}$$

However, as the ANC coefficient of the baseline parameter estimates the correlation between pre- and post- values the total variance is adjusted for intra-subject correlation as the intra-subject correlation forms part of the model. The more strongly correlated Y_0 and Y_1 are, the greater the reduction in the error variance. The residual variance in the ANC is given by [381]:

$$\mathrm{R}_i = [\sigma_1^2 - (\rho_{10})^2] \tag{11.4.10}$$

Table 11.4.5 shows the change in the variance of a pre-post test score across a range of correlation values. It can be seen from the table that:

$$V_i = u_{11}$$

Table 11.4.5 V_i matrix of the ANC method for a pre-post design

Correlation - $\boldsymbol{\rho}$	$\mathbf{R_i} = \left[\mathbf{1} - \boldsymbol{\rho}_{\mathbf{01}}^{\mathbf{2}}\right]$	$\boldsymbol{V_i} = (\mathbf{R_i})^{-\mathbf{1}}$	$\boldsymbol{u_{11}}$
$\mathbf{0}$	1	1	1
$\mathbf{0.25}$	0.9375	1.067	1.067
$\mathbf{0.50}$	0.75	1.333	1.333
$\mathbf{0.75}$	0.4375	2.286	2.286
$\mathbf{0.99}$	0.0199	50.25	50.25

Table 11.4.6 summarises the inverse covariance matrix used for each baseline analysis method in terms of the properties defined in Table 11.4.3.

Table 11.4.6 Inverse covariance matrix - V

Analysis strategy	**Inverse covariance matrix - $\boldsymbol{V}$**
RMA	$\begin{bmatrix} u_{00} & u_{01} & 0 & 0 \\ u_{10} & u_{11} & & \\ 0 & & \ddots & 0 \\ 0 & & 0 & \begin{matrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{matrix} \end{bmatrix}_{(N\cdot 2)\times(N\cdot 2)}$
ANC	$\begin{bmatrix} u_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & u_{11} \end{bmatrix}_{N\times N}$
CSA	$\begin{bmatrix} \frac{u_{11}}{w} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{u_{11}}{w} \end{bmatrix}_{N\times N}$

11.4.3.3 Comparison of $SE(\beta)$ – pre-post design

The standard error of the effect size estimates for each analysis method (RMA, ANC, CSA) can be derived from the $(X^TVX)^{-1}$ matrix for each analysis method, with full details presented in Appendix H. For simplicity, I present only the terms corresponding to the $treatment$ coefficient for the CSA and ANC methods and the $treatment \times time$ coefficient from the RMA (as these are the terms that capture the 'change' from baseline in each model). It must be noted that for the RMA and CSA models it is possible to derive generic $(X^TVX)^{-1}$ matrices that are applicable

to all cases however for the ANC assumptions must be made regarding the baseline imbalance or the $(X^TVX)^{-1}$ has a zero determinant and is non-invertible.

Table 11.4.7 $SE(\beta)$ for pre-post designs

Analysis Strategy	Model Parameter	$\boldsymbol{SE(\beta)}$
RMA	$treatment \times time$	$\sqrt{\frac{8}{N(u_{11}-u_{01})}}$
ANC	$treatment$	$\sqrt{\frac{4}{Nu_{11}}}$
CSA	$treatment$	$\sqrt{\frac{4}{N\frac{u_{11}}{w}}}$

Using the formulae for the $SE(\beta)$ from Table 11.4.7 and the values from Table 11.4.3 it is possible to calculate expected values for the standard error of the model coefficients. The results of these calculations are presented in Table 11.4.8. The table shows that for a pre-post design, the $SE(\beta)$ is identical for the CSA and RMA analyses across all correlation levels. By contrast, when the correlation strength between pre- and post- measurements is less than 1.0, the $SE(\beta)$ for the ANC is smaller than that of the CSA and RMA methods.

Table 11.4.8 Theoretical $SE(\beta)$ by analysis strategy – pre-post

Correlation	$\sum_{l,c=0}^{1} u_{lc} = w$	u_{11}	$\frac{u_{11}}{w}$	Standard Error of the Coefficient $(SE(\beta))$		
				RMA	ANC	CSA
0	2	1	$\frac{1}{2}$	0.20	0.1414	0.20
0.25	1.6	1.067	$\frac{2}{3}$	0.1732	0.1369	0.1732
0.5	1.3333	1.333	1	0.1414	0.1225	0.1414
0.75	1.142	2.286	2	0.10	0.0935	0.10
0.99	1.005	50.25	50	0.02	0.0200	0.02

11.5 Multi-follow-up design

A multiple follow-up design is a repeated measures experiment in which there is a single baseline measurement and follow-up is assessed more than once (two or more times). As with the pre-

post design, the primary efficacy measure is often the pattern or trajectory of change following treatment administration. The purpose of using multi-follow-up design is to (a) provide continuous monitoring of patients for unwanted side effect to the treatment and (b) decrease intra-patient variability and thus increase statistical power [382].

As compared with a pre-post design there are three distinct analysis complications introduced by having more than one follow-up assessment: 1) model parameter changes, 2) covariance structure and 3) mathematical coupling.

11.5.1 Model parameter changes

In moving from a pre-post design to a multi-follow-up, the CSA and ANC models will now have $time$ and t$reatment \times time$ terms.

The design and response matrices are now as follows:

Table 11.5.1 Multi-follow-up design and response matrices

Analysis Strategy	Design Matrix - $\boldsymbol{X}$	Response Vector - $\boldsymbol{Y}$
RMA	$\begin{bmatrix} 1 & x_{11} & x_{20} & x_{11}x_{20} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{11} & x_{2\mathrm{T}} & x_{11}x_{2\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1\mathrm{N}} & x_{2\mathrm{T}} & x_{1\mathrm{N}}x_{2\mathrm{T}} \end{bmatrix}_{(N\cdot m)\times 4}$	$\begin{bmatrix} y_{10} \\ \vdots \\ y_{1\mathrm{T}} \\ \vdots \\ y_{NT} \end{bmatrix}_{(N\cdot m)\times 1}$
ANC	$\begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} & y_{10} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{11} & x_{2\mathrm{T}} & x_{11}x_{2\mathrm{T}} & y_{1\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1\mathrm{N}} & x_{2\mathrm{T}} & x_{1\mathrm{N}}x_{2\mathrm{T}} & y_{\mathrm{NT}} \end{bmatrix}_{(N\cdot T)\times 5}$	$\begin{bmatrix} y_{11} \\ \vdots \\ y_{\mathrm{iT}} \\ \vdots \\ y_{NT} \end{bmatrix}_{(N\cdot T)\times 1}$
CSA	$\begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{11} & x_{2\mathrm{T}} & x_{11}x_{2\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1\mathrm{N}} & x_{2\mathrm{T}} & x_{1\mathrm{N}}x_{2\mathrm{T}} \end{bmatrix}_{(N\cdot m-1)\times 4}$	$\begin{bmatrix} y_{11} - y_{10} \\ \vdots \\ y_{\mathrm{NT}} - y_{\mathrm{N0}} \end{bmatrix} = \begin{bmatrix} \mathrm{D}_{11} \\ \vdots \\ \mathrm{D}_{NT} \end{bmatrix}_{(N\cdot T)\times 1}$

11.5.2 Covariance structure

As previously introduced in section 11.4.2, the residuals consist of a matrix of values where the diagonals of the matrix are the residual variances at each time point and the off diagonals are the covariances between successive time points:

$$R_i = Var(\varepsilon) = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0\mathrm{T}} \\ \sigma_{10} & \sigma_1^2 & \cdots & \sigma_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T0} & \sigma_{T1} & \cdots & \sigma_T^2 \end{bmatrix}$$

The covariance structure refers to the pattern in covariance matrices. Covariance structure is not relevant for pre-post analyses but is an important component of all multi-follow-up study analyses.

Three commonly used covariance structures previously introduced in section 0 are the CS, UN and AR(1) structures. Table 11.5.2 summarises the covariance for the $(k, l)^{th}$ element within each structure.

Table 11.5.2 Common covariance structures

Structure	Description	Parameters	$(\boldsymbol{kl})^{\boldsymbol{th}}$ element
CS	Compound Symmetry	2	σ_1 *or* $\sigma^2_{(k=l)}$
AR(1)	Autoregressive(1)	2	$\sigma^2\rho^{\lvert k-l\rvert}$
UN	Unstructured	$m(m+1)/2$	σ_{kl}

11.5.3 Mathematical coupling

Within the context of linear regression, mathematical coupling refers to a change in the correlation strength and structure of change scores. Mathematical coupling occurs when one variable directly or indirectly contains the whole or part of another [383]. As the baseline measurement is subtracted from all subsequent measurements, and moreover the baseline measurement forms a larger part of response measurements closer to baseline than those farther away, then the correlation between change scores will be different from that of the raw response scores. The correlation between a change score and the raw baseline measurement is given by [384]:

$$Corr(y_k - y_0, y_0) = \rho_{k-0,0} = \frac{\rho_{0k}\sigma_k\sigma_0 - \sigma_0^2}{\sqrt{\sigma_0^2 + \sigma_k^2 - 2\rho_{0k}\sigma_0\sigma_k}} \qquad (11.5.1)$$

Equation (11.5.1) has been expanded to accommodate the correlation between two change scores by [302]:

$$Corr(y_l - y_0, y_k - y_0 | k \neq l) = \frac{\rho_{kl}\sigma_k\sigma_l - \rho_{k0}\sigma_k\sigma_0 - \rho_{l0}\sigma_l\sigma_0 + \sigma_0^2}{\sqrt{\sigma_0^2 + \sigma_l^2 - 2\rho_{0l}\sigma_0\sigma_l}\sqrt{\sigma_0^2 + \sigma_k^2 - 2\rho_{0k}\sigma_0\sigma_k}} \quad (11.5.2)$$

As shown previously in section 0, if the original data structure is compound symmetry i.e. $\rho_{kl} = \rho_{l0} = \rho_{k0}$ then:

$$\rho_{k-0,l-0} = \frac{1-\rho}{2-2\rho} = \frac{1(1-\rho)}{2(1-\rho)} = \frac{1}{2} \quad (11.5.3)$$

If the original data structure was autoregressive then the change in correlation is slightly harder to explain. Consider a hypothetical study with 4 assessment time points, an AR(1) correlation structure and a correlation strength of 0.25 between the baseline and last assessment time point (see Table 11.5.3). The correlation matrix of time points is symmetrical around the diagonal, meaning that the correlation between any two time points is defined only by the distance between time points (i.e. the correlation between $k = 0$ and $k = 1$ is the same as the correlation between $k = 1$ and $k = 2$) – for this example, for a distance of 1 assessment the correlation is 0.63.

Table 11.5.3 AR(1) correlation structure for 4 time points

	Y_0	Y_1	Y_2	Y_3
Y_0	**1.0**	**0.63**	**0.40**	**0.25**
Y_1	**0.63**	**1.0**	**0.63**	**0.40**
Y_2	**0.40**	**0.63**	**1.0**	**0.63**
Y_3	**0.25**	**0.40**	**0.63**	**1.0**

Table 11.5.4 presents the correlation matrix of the correlation between the change scores D_1, D_2, and D_3 (where $=D_1 = Y_1 - Y_0, D_2 = Y_2 - Y_0, D_3 = Y_3 - Y_0$) calculated using equation (11.5.2). In contrast to the raw scores, the correlation is no longer symmetrical around the diagonal. The table shows that the correlation between D_1 and D_2 (0.64) is not equal to the correlation between D_2 and D_3 (0.73). The implication of this is that the correlation structure of the change scores is no longer AR(1).

Table 11.5.4 Correlation structure of the change scores from raw data with AR(1) structure

	D_1	D_2	D_3
D_1	**1.0**	**0.64**	**0.49**
D_2	**0.64**	**1.0**	**0.73**
D_3	**0.49**	**0.73**	**1.0**

Covariance matrix - The effect of coupling combined with the interaction between correlation structure and changes in variance associated with the baseline value in the ANC make a generic formulation of the covariance matrices complicated. Instead, we will illustrate these effects through a hypothetical study with four assessment time points, a correlation strength of 0.25 between the first and last assessment time point and a compound symmetry correlation structure (we have chosen compound symmetry as the computation of the residual variance is easier to follow).

Here we revert to the notation introduced in section 11.4.3.2 where we consider each covariance matrix to have the same generic form:

$$R_i = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \cdots & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 + \sigma_e^2 \end{bmatrix} \tag{11.5.4}$$

Where:

$$\begin{gathered} \sigma^2 = \sigma_a^2 + \sigma_e^2 \\ \sigma_a^2 = \rho \\ \sigma_e^2 = \sigma^2 - \sigma_a^2 = \sigma^2 - \rho \end{gathered} \tag{11.5.5}$$

Equation (11.5.4) shows that the R_i matrix contains only two specific values, the value of the diagonal elements and the value of the off diagonal elements; consequently the inverse of the R_i matrix ($R_i{}^{-1} = V_i$) will also have only 2 specific values. Note however that the RMA covariance matrix will be of dimension 4x4 whereas the ANC and CSA covariance matrices are of dimension 3x3. After some matrix algebra, formulae for the diagonal and off diagonal elements of $(R_i)^{-1}$ for each baseline analysis method can be obtained (Table 11.5.5).

Table 11.5.5 Diagonal and off-diagonal element values in the $(R_i)^{-1}$ matrices

	RMA	**ANC**	**CSA**
Diagonal	$\frac{3\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 4\sigma_a^2\sigma_e^2}$	$\frac{2\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2}$	$\frac{2\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2}$
Off Diagonal	$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 4\sigma_a^2\sigma_e^2}$	$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2}$	$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2}$

If we let σ^2(the total variance at each time point) be equal to 1 then the calculation for diagonal and off diagonal elements of the inverse covariance matrix for each analysis method are as follows:

RMA

$$\begin{gathered}\sigma^2 = \sigma_a^2 + \sigma_e^2 = 1 \\ \sigma_e^2 = 1 - \rho = 0.75 \\ \sigma_a^2 = \sigma^2 - \sigma_e^2 = 0.25\end{gathered} \tag{11.5.6}$$

The diagonal elements of R_i^{-1} will be equal to:

$$\frac{3\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 4\sigma_a^2\sigma_e^2} = \frac{3 \times 0.25 + 0.75}{0.75^2 + 4 \times 0.25 \times 0.75} = \frac{1.5}{0.75^2 + 4 \times 0.25 \times 0.75} = 1.429$$

And the off diagonal elements will be:

$$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 4\sigma_a^2\sigma_e^2} = \frac{-0.25}{0.75^2 + 4 \times 0.25 \times 0.75} = -0.1905$$

ANC

The CS variance is reduced by the baseline response being a covariate proportional to the correlation therefore:

$$\begin{gathered}\sigma^2 = 1 - \rho^2 = 1 - 0.25^2 = 0.9375 = \sigma_e^2 + \sigma_a^2 \\ \sigma_e^2 = 1 - \rho = 1 - 0.25 = 0.75 \\ \sigma_a^2 = \sigma^2 - \sigma_e^2 = 0.9375 - 0.75 = 0.1875 = \rho - \rho^2\end{gathered} \tag{11.5.7}$$

The diagonal elements of R_i^{-1} will be equal to:

$$\frac{2\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2} = \frac{2 \times 0.1875 + 0.75}{0.75^2 + 3 \times 0.1875 \times 0.75} = 1.429$$

And the off diagonal elements will be:

$$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2} = \frac{-0.1875}{0.75^2 + 3 \times 0.1875 \times 0.75} = -0.1905$$

CSA

The CSA covariance matrix is subject to regression to the mean, consequently, as introduced in section 7.3.1.2, the total variance will be inflated by the degree of correlation in the raw data, therefore:

$$\sigma^2 = 2(1-\rho) = 1.5 = \sigma_e^2 + \sigma_a^2$$
$$\sigma_e^2 = 1 - \rho = 1 - 0.25 = 0.75 \quad (11.5.8)$$
$$\sigma_a^2 = 1.5 - 0.75 = 0.75$$

And the diagonal elements of R_i^{-1} will be equal to:

$$\frac{2\sigma_a^2 + \sigma_e^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2} = \frac{2 \times 0.75 + 0.75}{0.75^2 + 3 \times 0.75 \times 0.75} = 1$$

And the off diagonal elements will be:

$$\frac{-\sigma_a^2}{(\sigma_e^2)^2 + 3\sigma_a^2\sigma_e^2} = \frac{-0.75}{0.75^2 + 3 \times 0.75 \times 0.75} = -0.333$$

Table 11.5.6 Summary of the calculated diagonal and off-diagonal elements of the $(R_i)^{-1}$ matrices

	RMA	**ANC**	**CSA**
Diagonal $(u_{kl}\|k = l) = u_{\text{diag}}$	1.429	1.429	1
Off Diagonal $(u_{\text{kl}}\|k \neq l) = u_{\text{off}}$	-0.1905	-0.1905	-0.333

We can now use these values in order to see how the standard error of the $treatment\ x\ time$ coefficient differs between baseline analysis methods.

For simplicity, I present only the terms corresponding to the $SE(\beta)$ for the $treatment\ x\ time$ interaction (as these are the terms that capture the 'change' from baseline in each model). It must be noted that the equations presented in Table 11.5.7 are applicable only to analysis scenarios in which there is compound symmetry between response time points.

Table 11.5.7 $SE(\beta)$ for treatment × time coefficient for multi-follow-up designs with compound symmetry correlation structure

Analysis Strategy	Model Parameter	$\boldsymbol{SE(\beta)}$
RMA	$treatment \times time$	$\sqrt{\frac{1}{(\frac{5N}{4})(u_{\text{diag}} - u_{\text{off}})}}$
ANC	$treatment \times time$	$\sqrt{\frac{1}{\frac{N}{2}(u_{\text{diag}} - u_{\text{off}})}}$
CSA	$treatment \times time$	$\sqrt{\frac{1}{\frac{N}{2}(u_{\text{diag}} - u_{\text{off}})}}$

Using the formulae for the $SE(\beta)$ from Table 11.5.7 and the values from Table 11.5.6 it is possible to calculate theoretical values for the standard error of the model coefficients. Assuming a sample size of 200 (100 per treatment arm), the results of these calculations are presented in Table 11.5.8. The table shows that for a multi-follow-up study with four patient assessment time points, the $SE(\beta)$ is identical for the CSA and ANC analyses across all correlation levels. By contrast, the $SE(\beta)$ for the RMA is smaller than that of the CSA and ANC methods across all correlation level. As the n is common to the denominator of all 3 calculations the relative difference between the analysis strategies will hold true across any sample size.

Table 11.5.8 Theoretical $SE(\beta)$ by analysis strategy – multi-follow-up with 4 patient assessments - compound symmetry correlation structure

Correlation	Standard Error of the Coefficient ($\boldsymbol{SE(\beta)}$)		
	RMA	ANC	CSA
0	0.063	0.100	0.100
0.25	0.055	0.087	0.087
0.5	0.045	0.071	0.071
0.75	0.032	0.050	0.050
0.99	0.006	0.010	0.010

11.6 Simulations

11.6.1 Design

Simulation studies were undertaken to assess the performance of each baseline analysis strategy under a variety of scenarios. From the theoretical calculations presented above, we defined the following four objectives:

1) Compare the theoretical calculations from Table 11.4.8 and Table 11.5.8 to empirical results from simulation
2) Compare the power of each baseline analysis method across number of time points where we hypothesise that when there is only 1 follow-up assessment (pre-post) the ANC will be more powerful than the RMA or CSA methods. By contrast when there is more than one follow-up assessment (multi-follow-up), the RMA will be more powerful than the ANC and CSA methods
3) As a consequence of mathematical coupling, we hypothesise that in multi-follow-up designs, CSA analyses specifying an unstructured covariance matrix will have an improved fit as compared with AR specification even when the raw response data has an AR correlation structure.

The power of each analysis method depends on the sample size, effect size, its standard error (as calculated above), and significance level (fixed at $\alpha = 0.05$). Effect size and sample size were chosen to give moderate power so that differences between the analysis methods can be readily observed (which is not possible if power is very low or very high).

Table 11.6.1 presents the power calculations using the Overall and Doyle (OD) Method from Equation (8.3.5 with the assumption of 100 patients in each treatment group. The table shows that an effect size (Δ) of 0.25 has power of between 46% and 84% across both correlation structures when the correlation is less than or equal to 0.5.

Table 11.6.1 Estimated power for various number of time points, effect size (Δ) and correlation strength and correlation structure assuming a two group comparison with 100 patients in each treatment group.

		OD Method							
Number of assessment time points (m)	**Δ**	**Autoregressive(1)**				**Compound Symmetry**			
		Correlation Strength $(\rho_{0,T})$							
		0	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**
4	0.1	0.11	0.13	0.17	0.29	0.11	0.14	0.18	0.32
	0.25	0.46	0.53	0.71	0.94	0.46	0.58	0.75	0.96
	0.5	0.96	0.98	1.00	1.00	0.96	0.99	1.00	1.00
6	0.1	0.13	0.13	0.17	0.29	0.13	0.16	0.22	0.39
	0.25	0.55	0.53	0.71	0.94	0.55	0.68	0.84	0.99
	0.5	0.99	0.98	1.00	1.00	0.99	1.00	1.00	1.00

Using the design number 6 data from Chapter 10 (no baseline imbalance, Δ=0.25) we sampled 100 patients in each study arm under two covariance structures (AR and CS), 5 correlation strengths (0, 0.25, 0.50, 0.75 and 0.99) and three different assessment time points (2, 4 and 6). For each scenario, 2000 simulated data sets were created, and each was analysed by the RMA, ANC and CSA methods. Each simulated data set was analysed using three different covariance structures, AR, CS, and unstructured (UN), regardless of the covariance structure used in the simulation.

11.6.2 Precision, power and fit

The precision of model coefficients $SE(\beta)$ is calculated from the standard deviation of the β coefficient across all 2000 simulations. The statistical power of each baseline analysis method was calculated as the percentage of simulations where the null hypothesis was rejected at a significance level of α=0.05. The fit of each analysis method was assessed using the Akaike information criterion (AIC).

11.6.3 Simulation results

11.6.3.1 Objective 1: Comparing theoretical values with results of simulation study

Table 11.6.2 presents the theoretical (from Table 11.4.8) and empirical (calculated from the simulation) calculations of the $SE(\beta)$ side by side. For pre-post studies (2 time points), the

simulation results confirm the theoretical expectation and show that the $SE(\beta)$ for the RMA and CSA methods is very similar across all levels of correlation. At low to intermediate correlation, the $SE(\beta)$ for the ANC is smaller than the $SE(\beta)$ for the RMA and CSA methods, however, the estimates converge at higher levels of correlation ($\rho > 0.75$).

Table 11.6.2 theoretical and empirical standard error of the coefficient ($SE(\beta)$) for pre-post study†–

Correlation	Theoretical			Empirical		
	RMA	**ANC**	**CSA**	**RMA**	**ANC**	**CSA**
0	0.20	0.14	0.20	0.2053	0.1420	0.2058
0.25	0.17	0.14	0.17	0.1723	0.1352	0.1723
0.5	0.14	0.12	0.14	0.1414	0.1211	0.1414
0.75	0.10	0.10	0.10	0.0952	0.0892	0.0952
0.99	0.02	0.02	0.02	0.0202	0.0201	0.0202

† Theoretical results were calculated using the methods presented in 11.4. Empirical results are derived from the standard deviation of the $\boldsymbol{\beta}$ across 2000 simulations each using 200 patients (100 per treatment arm).

For four time points, the theoretical (from Table 11.5.8) and empirical (calculated from the simulation) estimates of the $SE(\beta)$ are presented in Table 11.6.3. The simulation results confirm the theoretical expectation and show that $SE(\beta)$ is lower for the RMA than for the CSA and ANC across all levels of correlation.

Table 11.6.3 Theoretical and empirical standard error of the coefficient ($SE(\beta)$) for multi-follow-up study with four assessment time points and compound symmetry correlation structure†–

Correlation	Theoretical			Empirical		
	RMA	**ANC**	**CSA**	**RMA**	**ANC**	**CSA**
0	0.063	0.100	0.100	0.057	0.095	0.095
0.25	0.055	0.086	0.086	0.051	0.083	0.083
0.5	0.045	0.071	0.071	0.044	0.070	0.070
0.75	0.032	0.050	0.050	0.035	0.053	0.053
0.99	0.006	0.010	0.010	0.007	0.011	0.011

† Theoretical results were calculated using the methods presented in 11.4. Empirical results are derived from the standard deviation of the $\boldsymbol{\beta}$ across 2000 simulations each using 200 patients (100 per treatment arm)

11.6.3.2 Objective 2: comparing power by number of assessment time points

Figure 11.6.1 the empirical power results for each data correlation structure against correlation panelled by the number of assessment time points and data structure. For the RMA and ANC analysis methods the correct specification is used in the analysis (that is the model specified AR(1) for the AR(1) data and CS for the CS data). For the CSA analyses, we used an unstructured covariance matrix for all multi-follow-up AR analyses.

In general, across all three analysis methods (RMA, CSA and ANC), we observe that power:

1) Increases with higher intra-patient correlation.
2) Is higher with compound symmetry correlation structure as compared with autoregressive-1 correlation structure.
3) Increases with the number of assessment time points.

These results confirm our hypothesis that the ANC is the most powerful method in pre-post studies and that the RMA is the most powerful method in multi-follow-up studies. The power difference between the RMA and ANC/CSA methods is greater with compound symmetry than with auto-regressive correlation structure. With six time points, under auto-regressive correlation structure, the power was similar for all analysis methods.

11.6.3.3 Objective 3: covariance structure assumptions for simulation and analysis model

To assess the fit of each model, we used the Akaike information criterion (AIC) for each simulation, with low AIC values indicating a better model fit. Figure 11.6.2 displays the distribution of AIC values across the replications for the multi-follow-up scenario with *m*=6 and $\rho = 0.5$. The lower row of figures indicates that when the correlation structure in the simulation study is compound symmetry, using the correct model specification (CS) yields the best model fit. When the data structure is autoregressive (AR) (upper row), correct specification results in the best fit for the RMA and ANC methods. By contrast, for the CSA method, when the data correlation structure is AR, a model specification of UN gives improved model fit.

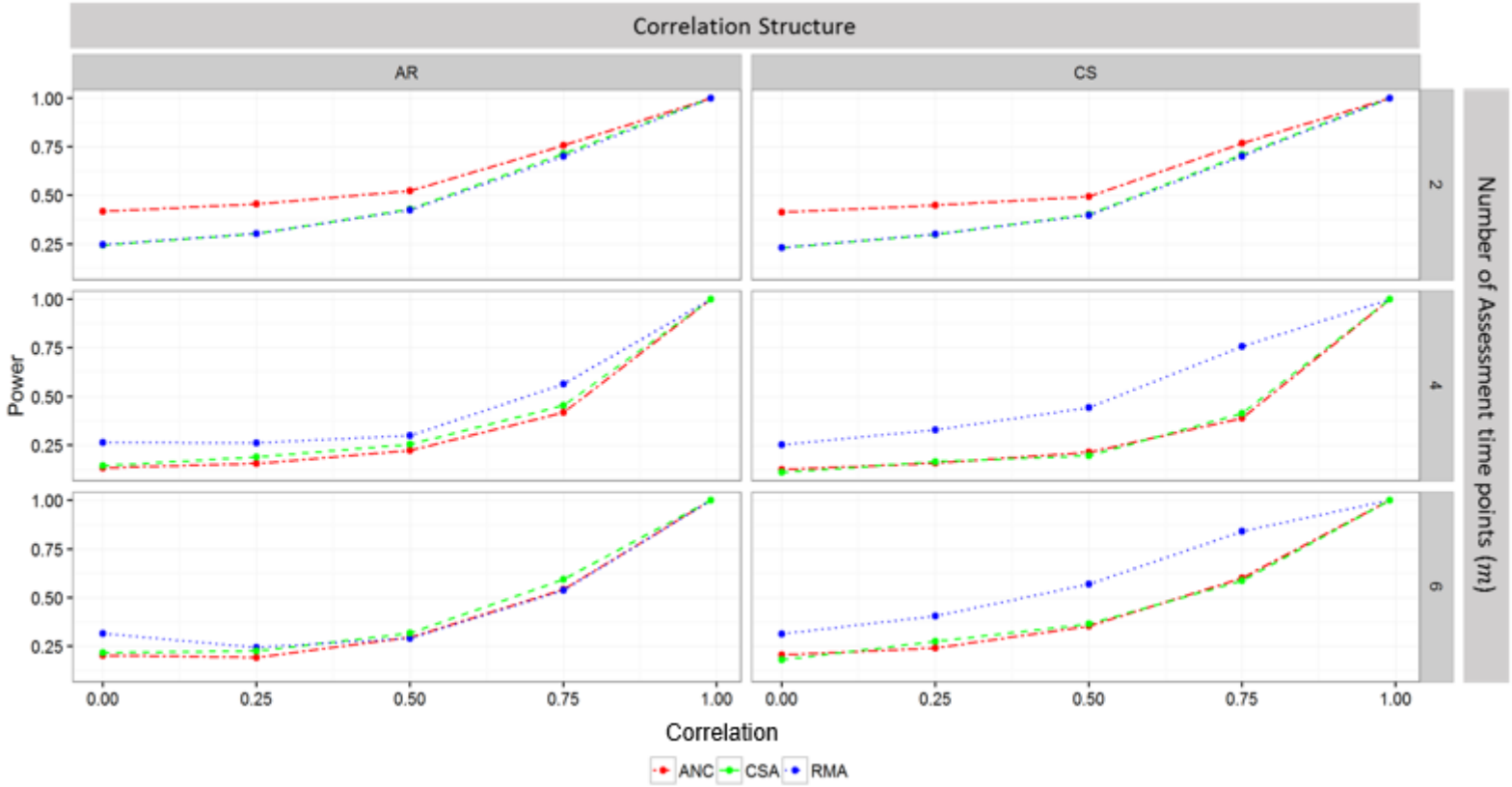


Figure 11.6.1 Power for each analysis method, by correlation structure and number of assessment time points.
For the ANC and CSA analyses, the model specification matched the data structure. For the CSA we used a correct specification for pre-post analyses and an unstructured covariance matrix for the multi-follow-up analyses. For pre-post studies, the ANC has more power than either the RMA or ANC methods. For multi-follow-up studies, RMA is more powerful.

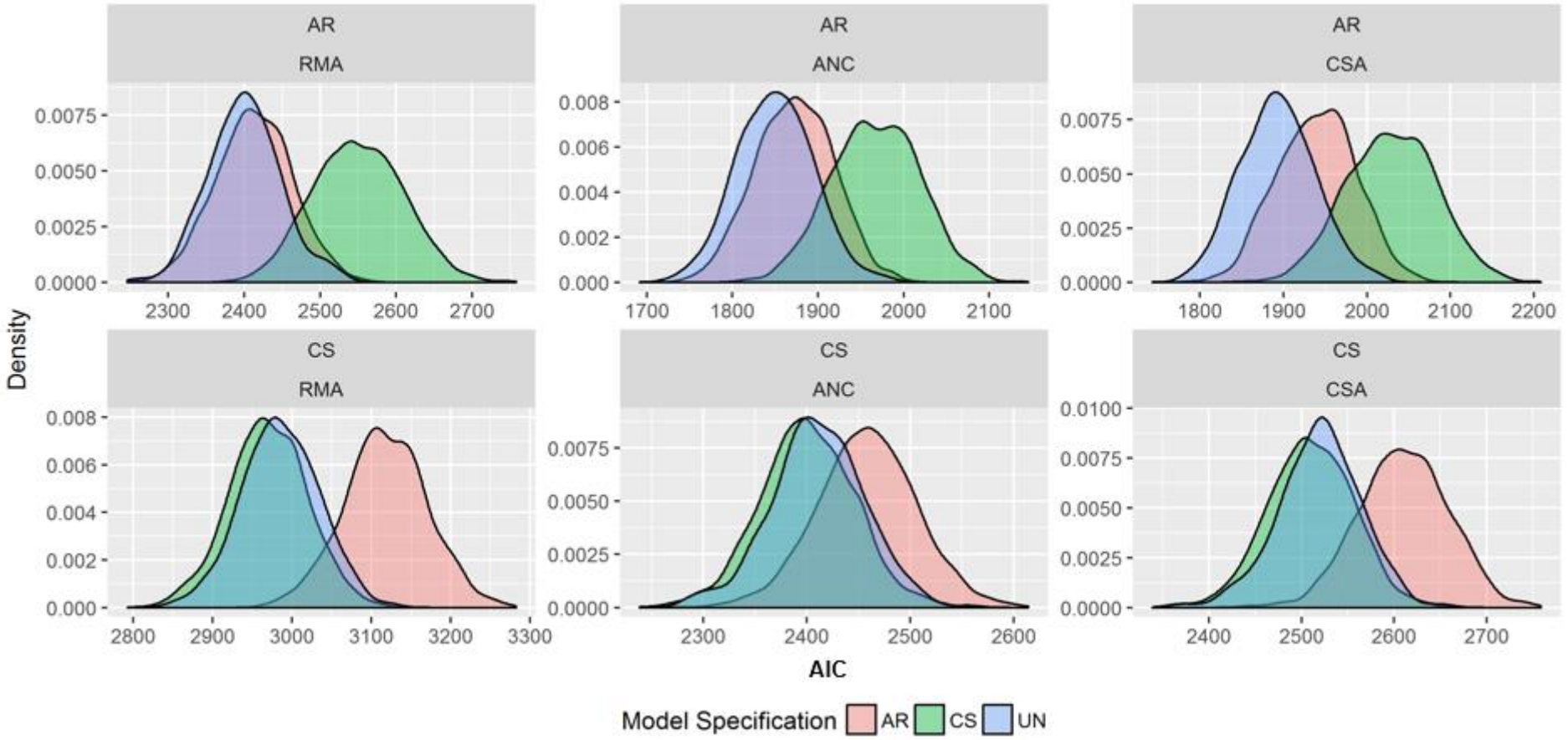


Figure 11.6.2 Density plot of the AIC for m=6 and correlation=0.5.
Each panel shows the fit of three model specifications for the covariance matrix (AR, CS and UN). For the CSA method, the UN covariance specification results in improved model fit when the data correlation structure is AR. For the RMA and ANC methods, model specification matching the correlation structure results in the best fit (smallest values), although in each case the unstructured method performs well.

11.7 Discussion

This study establishes the distinction between pre-post and multi-follow-up designs as a key factor in the choice of analysis for repeated measures clinical trials. Although the calculation and interpretation of the model coefficients from RMA, CSA and ANC have been described previously we are the first to examine the differences in the standard error of the coefficients in relation to the number of follow-up assessments, level of correlation and correlation structure.

Our results suggest 3 important findings:

1) When there is only a baseline and one follow-up assessment (pre-post) design, the ANCOVA (ANC) is the most statistically efficient method in the absence of baseline imbalance.
2) With more assessment time points (multi-follow-up), including the baseline measurement in the outcome vector (RMA) offers more power to detect a group x time interaction than either the ANC or CSA methods.
3) For multi-follow-up studies, the use of change scores (CSA) to analyse data that does not have compound symmetry correlation structure should specify an unstructured covariance matrix in the analysis.

This first finding supports the body of evidence where ANCOVA has been shown to be the analysis method of choice for incorporating baseline into the analysis of repeated measures data insofar as the analysis of pre-post design data. Vickers et al. [297], Senn et al. [385] and Egbewale et al. [87] all state that the ANCOVA has greater statistical power as compared with change scores, or using the baseline measure in the response vector. This improvement in power is commonly attributed to the adjustment in variance for the correlation between pre and post scores however we demonstrate that even in the absence of correlation the ANCOVA is more powerful than the RMA and CSA. This demonstrates that the gain in power for the ANC arises from differences in the variance but also is a consequence of a difference in the design matrix rather than ‘adjustment’.

If the trial collects more than a single follow-up assessment (multi-follow-up), including the baseline measurement in the outcome vector (RMA) is a more powerful technique than either the ANC or CSA. Both the CSA and ANC methods have a different number of predictors in multi-follow-up studies as compared with pre-post designs and consequently the design matrix for both these analysis methods changes when switching from a pre-post design to a multi-follow-up. We

observed that the critical difference between the RMA and CSA/ANC methods in multi-follow-up, which drives the disparity in power, is the additional baseline measurement in the response vector. While the specific difference between diagonal and off diagonal elements of the inverse covariance matrix is similar between analysis methods within a given correlation structure, the inclusion of the baseline in the response vector gives the covariance matrix of the RMA an extra row and extra column (corresponding to the extra time point). The consequence of this is that the $SE(\beta)$ for the $treatment\ x\ time$ coefficient in the RMA will almost always be smaller than the $SE(\beta)$ for the ANC and CSA methods, even when the covariance matrices hold the same values.

Because CSA uses outcome variables that derive from the subtraction each patient's baseline measurement, it has two distinct statistical challenges: 1) Regression to the mean and 2) mathematical coupling. Regression to the mean renders the CSA baseline analysis method statistically inefficient at low correlation for both pre-post and multi-follow-up designs. By contrast, mathematical coupling occurs only in multi-follow-up designs and is not recognisable from its effect on power. This is because statistical software requires user defined covariance matrices and misspecification of the covariance structure can result in an under or over estimation of the covariance between time points. Without specifically checking the model fit between covariance structures this error would go undetected and has the potential to inflate type-I error. It is clear from our results that when a change score is used for multi-follow-up data that is AR(1) (i.e. data that is not CS structure), the resulting covariance matrix will be asymmetrical and therefore an unstructured covariance matrix should always yield the best model fit. This is not to advocate the use of unstructured covariance matrices as the default option. The unstructured covariance matrix uses up a large number of degrees of freedom and would decrease power if a 'structured' covariance matrix would fit the data.

This work has several important limitations. Firstly, we fixed the variance to be equal across all assessment time points. This assumption may not be valid as heterogeneity of patient responses to treatment will likely become more prevalent over time [339]. Given that the covariance matrix is adjusted for the variance at each time the results of our baseline method comparison would still be expected to hold true even under simulated conditions of variance heterogeneity, however, future work might seek to confirm this. Similarly, we have assumed equally spaced assessment time points.

A second important point is that we assumed that the slope difference between baseline and the first post-baseline assessment was similar to the response profile of change over time for the post-baseline assessments. It is conceivable that the immediate/short-term response of patients could differ from the long-term response profile. If the initial response does not match subsequent response slope (and the on-treatment response slope is consistent) then the decision to include baseline in the outcome vector will result in a bias of the $treatment \times time$ interaction coefficient. Under these conditions, the exclusion of the baseline value by both the ANC or CSA analysis methods would be advantageous, regardless of the number of follow-up assessments. Even if the group response slope was similar, it is conceivable that the correlation between on-treatment responses might differ from the correlation between baseline and first on treatment assessment.

In conclusion, ANCOVA should be the analysis of choice for pre-post trials. The increase in power it offers over other baseline analysis methods is particularly marked when the correlation between pre- and post- measurements is low. In the absence of a discrepancy between the baseline and first post treatment response, and subsequent follow-up treatment responses, we believe that including the baseline in the response vector (RMA) should be the analysis method of choice for multi-follow-up studies. We specifically caution against the use of change scores in multi-follow-up studies as the correlation structure of the change scores may be altered from the correlation structure of the raw data, however, we note that misspecification of the correlation structure is potentially a problem for all analysis methods, and advocate that covariance assessment and covariance specification in repeated measures analyses become standard items of clinical study reports and publications from multi-follow-up studies.

Chapter 12. Discussion

12.1 Personalised medicine

Personalised medicine is a key goal of modern oncology. To achieve this vision of patient care researchers first need to identify the myriad factors - baseline, genetic and environmental, that shape a person's individual response to a particular treatment. This thesis used a series of studies to explore the link between efficacy and safety responses by identifying the factors that influence each. The data used in this thesis represent a unique opportunity to explore clinical and germline genetic factors for their association with platinum drug response. Germline genetic variation are associated with the pharmacokinetics of a drug [386, 387] and consequently may identify patients with increased risk of developing severe adverse events that could lead to treatment discontinuation or disruption. As both cohorts (lung and ovarian cancer) in this study received platinum containing regimens, the original study design envisaged using the cohorts for both discovery and replication of any polymorphisms associated with toxicity. Our work was focused on examining the similarities and differences between predictors of efficacy and predictors of toxicity, without the objective of replication. We explored the potential to combine both cohorts for identification of genetic and clinical factors that associate with response. In **Chapter 2** we observed differences between the cohorts with respect to baseline factors (i.e. gender, performance status, platinum subtype) that are known to influence the safety profile. Additionally, there were inherent differences between each cancer type with respect to the overall survival and progression free-survival curves. We concluded that the differences in baseline characteristics, safety profile and efficacy curves between cohorts prohibited the use of aggregate data (combining both cohorts) for the discovery of response associated factors.

In **Chapter 3** we introduced a framework for the link between efficacy and toxicity before exploring if toxicity responses during treatment predict 2-year efficacy status. In contrast to previous studies, the results of our analysis did not uncover an association between toxicity and efficacy responses. We believe that this may in part be due to the use of colony stimulating factor and antiemetic concomitant medications, both of which would confound the reporting of adverse events. Concomitant medication related exclusion criteria are among the most common barriers to enrolment in clinical trials. Since our data were from a treatment centre, rather than a clinical

trial, no such restrictions existed. Concomitant medication use in cancer patients is both common and varied and therefore it was not collected as part of the original study owing to the time commitment required to collect data retrospectively from patient medical files

Chapter 4 builds upon previous work that has examined baseline factors for their potential to predict improved efficacy outcomes and/or risk of experiencing severe adverse events. Our results confirm performance status as a significant independent predictor of PFS and OS efficacy endpoints in both lung and ovarian cohorts. We also observed cohort differences in prognostic factors. Performance status was prognostic of toxicity in the ovarian cohort while age and sex were prognostic of toxicity in the lung cohort. Sex was also an additional independent predictor for OS in the lung cohort and renal function was an additional predictor of OS in the ovarian cohort. Smoking status at baseline predicted for PFS in the lung cohort only. The difference between cohorts with respect to significant prognostic factors may reflect baseline differences in the cohort patient populations but may also be a reflect of the power differences between cohorts as the effective sample size (number of events) was higher in the lung cohort as compared with the ovarian cohort.

Pharmacogenomics have been widely recognised as fundamental steps toward personalised medicine. In **Chapter 5** we explored how germline variants are associated with patient response in each cohort. While this study identified two novel variants associated with overall survival in the lung cohort (rs17117678 and rs35075952), we noted that the addition of baseline covariates did not improve our ability to detect SNP-phenotype associations, as expected in a logistic regression analysis framework.

Lastly, in **Chapter 6** we developed a novel graphical method to describe the similarities of association results between the different clinical endpoints measured in cancer studies. Patients undergoing platinum chemotherapy typically experience several adverse reactions. While the toxicity phenotypes often co-occur, little is known about genetic correlation between adverse events. Our method characterises the similarities between association results of different phenotypes. If observed, this similarity could represent genetic correlation as a result of pleiotropy. It is hoped that with further development this method could help shed light on the

shared genetic basis of efficacy and toxicity phenotypes, and ultimately contribute to the understanding of the causal relationships among cancer therapeutic responses.

12.1.1 Limitations

Our work has several important limitations that must be discussed; these limitations can be broadly categorised into 1) sample size, 2) dosing and 3) factor independence. We discuss each below.

12.1.1.1 Sample size

A sample size with sufficient statistical power is important to the success of genetic association studies aiming to detect causal factors of human complex disease. In contrast to clinical covariates, genome-wide association studies require much larger sample sizes to achieve an adequate statistical power owing to the required alpha correction when testing a large number of potentially associated factors (SNPs).

The required sample size for detecting associations between disease and SNP markers is known to be a function of the following:

1) inheritance model (e.g. additive, dominant, or recessive)
2) prevalence (of response)
3) frequency of causal variant
4) coverage
5) effect size of the genetic variants (in our case - odds ratio or hazard ratio)

Genetic association tests for a relationship between the outcome (safety, efficacy, disease) and the genotype at a genetic variant. This is usually performed through a logistic regression, with SNP genotype as an explanatory variable, and including ancestry-informative principal components to control for population stratification. A SNP has three genotypes (e.g. AA, AB, BB) to be included in the model.

The genotypes for a SNP can also be grouped into inheritance models, such as dominant, recessive, or additive models [388, 389]. It is standard practice for genetic association studies to examine allelic models only, an approach that we adopted for our analyses. Previous work [389, 390] has indicated the allelic model as the standard analysis approach as it has reasonable power to detect both additive and dominant effects. If however the true underlying model of inheritance

was recessive then this assumption would result in underpowered analyses to detect SNP-phenotype associations [389]. Rather than choosing one model *a priori*, we might have opted to evaluate multiple genetic models. This would need to have concomitant alpha correction for multiple testing across all genetic models thereby reducing our power to detect significant SNP-phenotype associations.

Prevalence is the proportion of individuals with the affected phenotype. Within our study the prevalence of adverse events varied by efficacy and toxicity phenotype. Of the recorded adverse event terms, only severe neutropenia was experienced by 10% of patients in both cohorts. Consequently to explore the effects of gastrointestinal intestinal upset we mapped adverse event terms to MedDRA system organ class (SOC) [391]. As not all SOC groupings are based upon aetiology, then an analysis of SOC terms might contain adverse event terms with disparate genetic underpinnings that reduce our power to detect SNP-'SOC' associations.

For survival (efficacy) endpoints the 'prevalence' is less applicable, as the survival analysis aims to answer the question regarding differences between levels of a prognostic factor in the time to affected status. If all patients achieved affected status by the end of the study period, then the effective sample size would match the number of patients. As the partial likelihood function is essentially based on the number of events the effective sample size is usually much lower than the number of patients within the trial [392]. In the lung cohort, the proportion of patients experiencing OS and PFS events was greater than in the ovarian cohort. Differences in the response prevalence between phenotypes (efficacy and safety outcomes) and cohorts, render comparison of association results difficult.

Illumina's HumanExome array targets around 250,000 SNPs. The marker SNPs were identified from an analysis of over 12,000 whole exome sequenced genomes individuals from diverse populations, including European, African, Chinese, and Hispanic individuals, and includes SNPs associated with a range of common conditions including type 2 diabetes, cancer, metabolic, and psychiatric disorders. The chip was designed to help researchers in identifying functionally relevant associations within exonic regions; and it has proven a useful tool in identifying such associations. However, the human exome is less than 1% of the genome (30 megabases, compared to 3.3 gigabases in the whole genome), and therefore this chip is not equivalent to a

genome-wide genotyping array where coverage is designed to capture common variation across the genome. Many SNPs on the Illumina HumanExome array have low frequency, and are not in linkage disequilibrium with other SNPs on the array. Our analyses identified two SNPs that were significantly associated with overall survival in lung cancer patients receiving platinum-based chemotherapy, one within the *OMA1* and the other within the *TACSTD2* gene. In both cases, none of the other tagged SNPs in the region was in strong enough linkage disequilibrium to observe supporting association signals for either significant SNP. Replication studies will be essential to determine whether these are true findings or false positive results, but at present, few similar studies exist.

12.1.1.2 Dosing

The data collected in this study did not include patient specific dosing information or patient body size. The specific quantity of platinum chemotherapy administered to each patient varies by body size. Platinum therapy is dosed according to body surface area (BSA) which is calculated as a function of patient height and weight [393, 394]. It has been suggested that BSA-based dosing results in therapeutic inconsistencies between patients as it does not account for the patient variations in hepatic or renal function, the proportion of muscle and fat tissues, or multiple other variables involved with drug processing [395]. These inconsistencies may lead to a variation in drug exposure and therefore response.

As inter-patient variation in the dose intensity by cycle will likely contribute to patient efficacy and safety responses, it represents a potential confounding factor when identifying prognostic variables.

Another potentially critical piece of information not contained within the data was patient specific records of concomitant medications. Common side effects of platinum drugs include emesis and neutropenia. When neutropenia is observed in a patient, granulocyte colony stimulating factor (G-CSF) may be prescribed to help stimulate white blood cell recovery. Similarly, severe or persistent emesis is often managed through antiemetic medications. Both of these concomitant medications would result in an underreporting of adverse event endpoints and reduce our power to identify prognostic factors associated with gastrointestinal and neutropenic outcomes.

12.1.1.3 Factor independence

Associations between cancer prognosis and common but low penetrance genetic polymorphisms may not be independent of other known clinical prognostic factors. In order to clarify the independent contribution of germline genetic polymorphisms beyond known clinical prognostic factors, it has been suggested that studies exploring clinical outcome associations with germline polymorphisms take into account clinical prognostic factors in addition to genetic variation [396]. Consequently, within our work, we explored the addition of the prognostic factors age and sex when identifying germline variant association with clinical outcomes. The addition of these baseline patient characteristics consistently weakened our SNP-outcome associations, and while we concluded that this was a feature of the statistical methods employed it is possible that the variants identified are not acting independently of the baseline characteristics.

Until we are able to assess the genetic contribution of germline variants to baseline features such as age of disease onset, performance status and renal function, the independent role of germline genetic variations upon patient responses will remain hard to delineate.

Clarifying this will require large-scale replication studies of germline polymorphisms and drug response that extend beyond hypothesis-generating studies that currently dominate medical literature.

12.1.2 Future directions

Future work on this dataset could revisit the patient records to collect drug dosing information, concomitant medications and more detailed body mass information. This information could be used to explore the body mass adjusted association between dose and the exposure-adjusted risk of adverse events. A refined understanding of the interplay between BMI, dose and total exposure could then be used to construct an improved model with which to detect SNP-AE associations, and could be used to update the analysis in which adverse events were used to predict efficacy outcomes. Once reliable SNP-outcome associations are established for both safety and efficacy endpoints, a scoring model, akin to a polygenic risk score, could combine a patient's variant information to provide SNP-informed benefit-risk assessment. For example, if the specific variants carried by an individual predict an above average risk for an adverse event, and a below average survival response, then the benefit risk might be considered unfavourable.

Balancing the efficacy and safety of prescribed chemotherapy treatments is crucial to optimising chemotherapy treatment outcomes. Individual variability in drug efficacy and safety remain a challenge in drug development and clinical practice. Knowledge of genetic determinants of efficacy and safety outcomes promises a potential pathway to achieving the ultimate goal of personalised medicine within oncology.

12.1.3 Conclusions

Every day, patients are exposed to medications that will have limited or no benefit to them. Classical clinical trials harvest a handful of measurements from thousands of people [397]. Such trials result in treatment decisions that follow a 'one size fits all' approach as clinical trials are typically unable to distinguish patient characteristics that interact with response within a given treatment arm. Personalised medicine is essentially the ability to tailor treatments, as well as prevention strategies, to the unique characteristics of each person. When it comes to cancer, personalisation can take several different forms. Advances in our ability to characterise genetic changes that occur within tumours has resulted in targeted therapeutics. Studies exploring the role of germline or inherited genetic variation on cancer outcome have identified germline polymorphisms associated with differential survival and toxicity. Lastly, numerous studies have identified baseline patient characteristics that can be used to stratify patient response.

This thesis has examined both patient characteristics and germline genetic factors for their association with efficacy and safety responses in patients receiving platinum therapy. It is hoped that these studies will contribute towards identification of the relevant prognostic factors that shape patient response to platinum therapy. Investigation of the potential use of germline genetic variation as prognostic and predictive markers is a relatively new area of research. Advances in predicting patient responses based on current, common therapeutic strategies represent a key step in the process of bringing personalised medicine to clinical practice

12.2 Repeated measures clinical trials

In the recent past, licencing approval of a therapeutic agent within oncology depended primarily on the demonstration of improved efficacy outcomes in patients treated with the new agent as compared with patients treated with the standard of care. Increasingly the strategy is evolving to

accept a novel compound if it can be demonstrated that there is similar efficacy with reduced toxicity. As a result of the increasing desire to minimise toxicities, adverse events are closely scrutinized and patients enrolled in clinical trials are intensively monitored for safety events that may relate to treatment. Consequently, oncology trials ubiquitously collect repeated measurements of adverse events during treatment and frequently even after treatment cessation. The wealth of safety data generated has led biostatisticians to question if the longitudinal safety data can be used to elucidate treatment effects on clinical outcomes such as overall survival or progression free survival. One popular methodological approach to this problem involves the joint modelling of survival with longitudinal data. In this approach, longitudinal mixed models are used to incorporate the effects of time-varying covariates in the evaluation of efficacy endpoints [398]. Outside of oncology, longitudinal values of blood pressure [399, 400] and CD4[3] cell counts [401-403] have previously been used in the evaluation of treatment influence upon disease progression and survival. At the onset of this thesis, one of the primary objectives was to use longitudinal information of adverse events to inform the survival process. To prepare for this we undertook a review of longitudinal marginal models, the results of which are presented in this thesis and summarised below. Upon examination of the adverse events across the four cycles of therapy, we discovered that the changes in CTCAE grade between cycles did not follow linear patterns, and therefore it would not be possible to model this data as originally planned. Our analyses of adverse events dichotomised patients based upon the adverse event severity at any time during therapy. Despite not being able to longitudinally model the adverse event data, our early examination of longitudinal methods revealed items relevant for the design and analysis of repeated measures studies. Given the prevalence of repeated measures trial designs, we believe that this information can contribute to the wider field even if it was not applicable to our data. Below we summarise our findings from the work on repeated measures linear regression in Chapters 7 to 11.

[3] Cluster of differentiation 4; a protein encoded by the *CD4* gene which is found on the surface of immune cells.

12.2.1 Repeated measures linear regression

The analysis of repeated measures data represents a continuing challenge for clinical trials collecting repeated measurements on each subject. While generalised linear models have emerged as the standard for analysing repeated measures data, there are several commonly employed strategies for handling the baseline measurement:

1) Retain it as part of the outcome vector (RMA);
2) Use the baseline measurement as a covariate in the analysis of the post-baseline measurements (ANC);
3) Subtract the baseline measurement from all the remaining post-baseline measurements and then analyse the change scores (CSA).

The relative strengths and weaknesses of each strategy have been extensively covered for pre-post designs, however, we found very few studies examining each analysis strategy where there is more than one follow-up assessment or across alternate correlation structures. Moreover, despite the frequent claim that the ANCOVA is more precise/efficient and therefore powerful in pre-post designs, we were not able to readily identify the exact reason for this and sought to elucidate.

In Chapter 7 we review how sample size is estimated for repeated measures experiments. This chapter introduces the idea that in longitudinal studies the natural experimental unit is no longer an individual measurement in time, but the vector of patient measurements that give rise to linear mean differences between treatment groups across time. It also introduces the critical concepts of a covariance matrix and correlation structure as key features in determining the required sample size to achieve $1-\beta$ in repeated measures experiments. Chapter 8 uses the formulae from Chapter 7 to present an analytic study exploring how the correlation strength, correlation structure, effect size and number of assessment time points influence the sample seize required to achieve a given level of power. This work reveals that for a linear trend with evenly spaced assessment time points, there is no gain in power associated with the transition from a pre-post design to a design with two follow-up assessments. This finding has major implications for trial design and efficiency of trial resources as it indicates that increasing the sample size may represent the best use of trial resources when the frequency of assessment cannot be increased beyond three patient assessment time points. We also confirm the previous finding that increased correlation strength reduced the sample size required when the correlation structure is compound

symmetry, but that low and intermediate correlation strength require an increased sample size when the correlation structure is autoregressive.

As information regarding intra-patient correlation strength and structure is often unavailable at the time of study planning, it is not uncommon for researchers to assume that the correlation between baseline and post-intervention measures is zero and calculate a sample size for a simple two-sample comparison of means. We show that this practice results in dramatic inflation of the required sample size. Further, we conclude that assuming an autoregressive correlation structure with low patient correlation results in adequate power across all correlation strengths and structures while significantly reducing the estimated sample size from the estimate produced by the comparison of group means at final assessment.

Chapter 9 reviews two alternate generalised linear model used in repeated measures data analysis: the marginal model and the mixed-effects model. This chapter also presents alternate baseline analysis strategies and reviews previous research that has examined alternate strategies. Given the frequency of repeated measures trials and the length of time that repeated measures analytical methods have been available, it was surprising to find that the statistical analysis of change is still mired in controversy over how to best handle the baseline measurement. Despite the wealth of literature, we still found an absence of clear guidance regarding the implications of alternate baseline analysis strategies upon the estimated model coefficients. Therefore, in Chapter 10 we use simulation across a range of hypothetical trial scenarios to ascertain how model coefficients relate to specific trial design parameters.

Our work demonstrates that the interpretation of the model estimates depends on the number of assessment time points and, that the choice between analysis methods should depend on whether the experiment is a pre-post design or has more than one follow-up measurement. This work is able to describe each model coefficient in relation to the intercepts and slopes of each treatment arm, and describes how the interpretation of model estimates changes when transitioning from a pre-post to a multi-follow-up design.

A key finding of this work is that the estimated coefficients for ‘change’ are equal between baseline analysis methods for pre-post designs in the absence of baseline imbalance, and equal irrespective of baseline imbalance for multi-follow-up trials (more than one follow-up assessment).

As power is a function of both the estimated coefficient and the standard error of the coefficient, we explore in Chapter 11 the similarities and differences between baseline analysis strategies for the calculation of the variance of the estimates. Through a combination of algebraic solution and simulation, we detail how each method differs in the covariance matrix used to estimate variance of the estimate, and explore the implications of these differences upon statistical power. Specifically, we show that for a pre-post design, the ANCOVA has the greatest power, but when there is more than one follow-up measurement, the greatest power is achieved by using the baseline measurement as part of the response vector. Due to mathematical coupling, we show that calculation of the change score can result in a non-symmetric covariance structure, and therefore advocate the use of an 'unstructured' covariance specification when the analysed data has an autoregressive correlation structure. Lastly, we show how the calculation of the standard error of the estimate changes for ANCOVA and change-score methods depending on the number of time points, which explains the difference in power between baseline analysis methods when moving from a pre-post design to a repeated measures experiment in which there is more than a single follow-up measurement.

12.2.2 Limitations

Our work has several important limitations that offer the potential for future directions in this line of research; these limitations can be broadly categorised into:

1) covariance structure misspecification
2) missing data
3) random effects

12.2.2.1 Covariance structure misspecification

Misspecification of the correlation structure within a repeated measures analysis refers to specifying a correlation structure within the analysis that does not match the correlation structure of the data being analysed. Previous researchers have identified [404] three categories of covariance structure misspecification: underspecified, overspecified and general misspecification. Underspecified refers to a model specified covariance structure that is simpler than the true covariance matrix (i.e. compound symmetry is specified in the model but the true structure is AR(1)). Overspecification refers to a model specified covariance structure that is more complex than the true covariance matrix with the caveat that the true covariance matrix is

nested within the specified matrix (e.g. ARMA(1) structure chosen but AR(1) is the true structure). Lastly, general misspecification occurs when the specified and true covariance matrices are not nested (i.e. CS structure chosen but AR(1) is the true structure). Simulation studies examining the consequences of covariance structure misspecification have found either no or modest bias in the estimates of fixed effects. There is evidence that the bias for the estimated standard errors of the fixed effects is much stronger [404-406]. These studies identified that both underspecification and general misspecification result in inflation of the estimated standard errors even in the absence of coefficient bias. Conversely, overspecification of the covariance structure resulted in smaller estimates of coefficient standard errors as compared to the correct specification. These articles did not explore the impact of the frequency of patient assessment or compare baseline analysis strategies. Given the observed differences between each baseline analysis method with respect to the covariance matrices, it is reasonable to assume that misspecification would result in differential variable inflation/deflation of the standard errors across methods.

12.2.2.2 Missing data

The work in this thesis assumed that all patients had complete data, a potentially unrealistic scenario. Data missingness is a common feature of repeated measures clinical trials that represents a major challenge during the analysis of the longitudinal data. Missing data can have both different patterns and causes, all of which have the potential to bias the estimates and/or efficiency of analysis results [407-410]. Future work could systematically examine if missingness results in bias and/or loss of power equally between each baseline analysis method across a range of alternate types and quantities of missing data.

12.2.2.3 Random effects

Chapter 9 introduced both marginal and mixed-effect generalised linear models for the analysis of repeated measures data. A marginal model adjusts for repeated measures through estimation of the covariance of residuals. By contrast, mixed models adjust for repeated measures by altering the model to include random effect (leaving the residuals alone). In recent years, two types of mixed models have merged as popular choices for repeated measures data. These are:

1) The random intercept model, and
2) The random slope model.

The addition of a random intercept term allows each patient to deviate from the overall mean response by a person-specific constant that applies equally over time. Similarly, the random slope model adds a random term that captures how the patient specific slope varies from the overall mean slope parameter for the study population. Mixed models have been shown to be robust to missing data and irregularly spaced measurement occasions [411]. As such, they are among the most popular methods for the analysis of longitudinal data. As our work did not incorporate missing data we chose to explore the alternate baseline analysis strategies using marginal models. Future work would seek to explore if there are similar differences between baseline analysis strategies with respect to the D covariance matrix of random effects as were observed with the R covariance matrix of residuals from the marginal model.

12.2.3 Conclusions

In summary, repeated measures regression models offer a robust method for the analysis of repeated measures. The ability to model and compare longitudinal response patterns can increase statistical power and therefore represent a major strength of this methodology. Due to the greater statistical power, a repeated measures design can use fewer subjects to detect a desired effect size and therefore represent an effective use of research resources. However, repeated measures designs require that care be taken to understand the correlation structure between repeated measures. The correlation structure gives rise to the covariance matrix, which is a critical component in the calculation of both the parameter estimates and their standard errors. While modern software will estimate a study specific covariance matrix, it is still necessary for the researcher to define the covariance structure that will be used in the estimation. We have shown that alternate methods for inclusion of the baseline measurement in the analysis of repeated measures data have important consequences for the covariance matrix that affect the power of each method. Therefore, we advocate that repeated measures studies should rountinely report information regarding the strength and structure of intra-patient correlation between assessment time points.

In summary, this thesis has performed research in two different areas of relevance to personalised medicine. The genetic study of efficacy and safely in cancer chemotherapeutics showed the

challenges integral to analysis of clinical cohorts. Patient heterogeneity is a substantial barrier to accruing the large sample sizes that are essential for powerful genetic association studies. Differences in cancer site and medication – as well as individual patient characteristics of age, and sex – all impact the choice of analyses to be performed. Our work developed a novel exploratory data analysis tool to explore any correlation between genetic associations for efficacy and safety.

Clinical studies and trials often produce longitudinal data with repeated measures on patients or study participants. The second part of this thesis considered in detailed the choice of analytical method for this study design. We showed that care is needed in how to deal with the baseline measurement, with the choice of method depending on the number of assessments for each patient, and the correlation structure between observations at different time points. Our work adds to the literature in this area by showing that ANCOVA methods are most powerful for pre-post studies, and we advise that for multi-follow-up studies, using the baseline measurement as an outcome variable increases power. Both these research areas illustrate the importance of rigorous statistical analysis in clinical studies. Personalised medicine will require multi-disciplinary clinical and research teams, with statisticians having a key role to ensure effective studies are performed, that maximise the potential for patient benefit.

References

1. Seegobin, S.D., et al. (2014). ACPA-positive and ACPA-negative rheumatoid arthritis differ in their requirements for combination DMARDs and corticosteroids: secondary analysis of a randomized controlled trial. *Arthritis Research & Therapy*. **16**(1): p. R13.
2. Opdam, F.L., H. Gelderblom, and H.J. Guchelaar. (2012). Phenotyping drug disposition in oncology. *Cancer Treatment Reviews*. **38**(6): p. 715-725.
3. Trusheim, M.R., E.R. Berndt, and F.L. Douglas. (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews: Drug Discovery*. **6**(4): p. 287-293.
4. Schilsky, R.L. (2010). Personalized medicine in oncology: the future is now. *Nature Reviews: Drug Discovery*. **9**(5): p. 363-366.
5. Oldenhuis, C.N., et al. (2008). Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*. **44**(7): p. 946-953.
6. Goldstraw, P., et al. (2007). The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *Journal of Thoracic Oncology*. **2**(8): p. 706-714.
7. Clark, T.G., et al. (2001). A prognostic model for ovarian cancer. *British Journal of Cancer*. **85**(7): p. 944-952.
8. Berman, M.L. (2003). Future directions in the surgical management of ovarian cancer. *Gynecologic Oncology*. **90**(2 Pt 2): p. S33-39.
9. Zheng, Q.-Q., et al. (2009). Prognostic analysis of ovarian cancer patients using the Cox regression model. *Chinese Journal of Cancer*. **28**(2): p. 170-172.
10. Sculier, J.P., et al. (2007). Chemotherapy improves low performance status lung cancer patients. *European Respiratory Journal*. **30**(6): p. 1186-1192.
11. Albain, K.S., et al. (1991). Survival determinants in extensive-stage non-small-cell lung cancer: the Southwest Oncology Group experience. *Journal of Clinical Oncology*. **9**(9): p. 1618-1626.
12. Takigawa, N., et al. (1996). Prognostic factors for patients with advanced non-small cell lung cancer: univariate and multivariate analyses including recursive partitioning and amalgamation. *Lung Cancer*. **15**(1): p. 67-77.
13. Umesaki, N., et al. (1993). Studies of prognostic factor and chemotherapeutic effect of epithelial ovarian cancer using Cox's proportional hazard model. *Gan to Kagaku Ryoho*. **20**(15): p. 2351-2356.
14. Palomares, M.R., et al. (1996). Gender influence on weight-loss pattern and survival of nonsmall cell lung carcinoma patients. *Cancer*. **78**(10): p. 2119-2126.
15. Bonomi, P., et al. (1991). Pre-treatment prognostic factors in stage III non-small cell lung cancer patients receiving combined modality treatment. *International Journal of Radiation Oncology, Biology, Physics*. **20**(2): p. 247-252.
16. Feinstein, A.R. and C.K. Wells. (1990). A clinical-severity staging system for patients with lung cancer. *Medicine (Baltimore)*. **69**(1): p. 1-33.
17. Thorogood, J., et al. (1992). The use of discriminant analysis to guide palliative treatment for lung cancer patients. *Clinical Oncology (Royal College of Radiologists)*. **4**(1): p. 22-26.
18. Massard, G., et al. (1996). Bronchogenic cancer in the elderly: operative risk and long-term prognosis. *Thoracic and Cardiovascular Surgeon*. **44**(1): p. 40-45.
19. Karim, N.A., et al. (2014). The use of pharmacogenomics for selection of therapy in non-small-cell lung cancer. *Clinical Medicine Insights: Oncology*. **8**: p. 139.
20. Bristow, R.E., et al. (2013). Disparities in ovarian cancer care quality and survival according to race and socioeconomic status. *Journal of the National Cancer Institute*. **105**(11): p. 823-832.
21. Siegel, R., D. Naishadham, and A. Jemal. (2013). Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*. **63**(1): p. 11-30.
22. Sobue, T., et al. (1991). Prognostic factors for surgically treated lung adenocarcinoma patients, with special reference to smoking habit. *Japanese Journal of Cancer Research*. **82**(1): p. 33-39.
23. Kelemen, L.E., et al. (2016). Smoking may modify the association between neoadjuvant chemotherapy and survival from ovarian cancer. *Gynecologic Oncology*. **140**(1): p. 124-130.
24. Kristyanto, H. and A.R. Utomo. (2010). Pharmacogenetic application in personalized cancer treatment. *Acta Medica Indonesiana*. **42**(2): p. 109-115.

25. Mcwhinney, S.R., R.M. Goldberg, and H.L. Mcleod. (2009). Platinum neurotoxicity pharmacogenetics. *Molecular Cancer Therapeutics*. **8**(1): p. 10-16.
26. Potti, A., et al. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*. **12**(11): p. 1294-1300.
27. Iyer, L., et al. (1999). Phenotype-genotype correlation of in vitro SN-38 (active metabolite of irinotecan) and bilirubin glucuronidation in human liver tissue with UGT1A1 promoter polymorphism. *Clinical Pharmacology and Therapeutics*. **65**(5): p. 576-582.
28. Pullarkat, S.T., et al. (2001). Thymidylate synthase gene polymorphism determines response and toxicity of 5-FU chemotherapy. *Pharmacogenomics Journal*. **1**(1): p. 65-70.
29. Salgado, J., et al. (2007). Polymorphisms in the thymidylate synthase and dihydropyrimidine dehydrogenase genes predict response and toxicity to capecitabine-raltitrexed in colorectal cancer. *Oncology Reports*. **17**(2): p. 325-328.
30. Rocha, V., et al. (2009). Association of drug metabolism gene polymorphisms with toxicities, graft-versus-host disease and survival after HLA-identical sibling hematopoietic stem cell transplantation for patients with leukemia. *Leukemia*. **23**(3): p. 545-556.
31. Nowell, S.A., et al. (2005). Association of genetic variation in tamoxifen-metabolizing enzymes with overall survival and recurrence of disease in breast cancer patients. *Breast Cancer Research and Treatment*. **91**(3): p. 249-258.
32. Amado, R.G., et al. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*. **26**(10): p. 1626-1634.
33. De Roock, W., et al. (2008). KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. *Annals of Oncology*. **19**(3): p. 508-515.
34. Apperley, J.F. (2007). Part I: mechanisms of resistance to imatinib in chronic myeloid leukaemia. *Lancet Oncology*. **8**(11): p. 1018-1029.
35. Abramson, R.G. (2016) *Overview of targeted therapies for cancer*. [cited 2016 07/11]; Available from: https://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/.
36. Rodriguez-Vida, A., M. Strijbos, and T. Hutson. (2016). Predictive and prognostic biomarkers of targeted agents and modern immunotherapy in renal cell carcinoma. *ESMO Open*. **1**(3): p. e000013.
37. Alsop, K., et al. (2012). BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. *Journal of Clinical Oncology*. **30**(21): p. 2654-2663.
38. Bolton, K.L., et al. (2012). Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA*. **307**(4): p. 382-390.
39. Sinicrope, F.A., et al. (2011). DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. *Journal of the National Cancer Institute*. **103**(21): p. 1639-1639.
40. Sinicrope, F.A., et al. (2013). Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy. *Journal of Clinical Oncology*. **31**(29): p. 3664-3672.
41. Geiger, T. (2014). Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy. *Diseases of the Colon and Rectum*. **57**(3): p. E39-E40.
42. Rottenberg, S., et al. (2008). High sensitivity of BRCA1-deficient mammary tumors to the PARP inhibitor AZD2281 alone and in combination with platinum drugs. *Proceedings of the National Academy of Sciences of the United States of America*. **105**(44): p. 17079-17084.
43. Ang, J.E., et al. (2013). Efficacy of chemotherapy in BRCA1/2 mutation carrier ovarian cancer in the setting of PARP inhibitor resistance: a multi-institutional study. *Clinical Cancer Research*. **19**(19): p. 5485-5493.
44. Chowdhury, K., et al. (2015). Combination treatment of PARP inhibitor, BMN 673 and DNMT inhibitor, Azacytidine: A potential therapy for BRCA negative and positive, triple negative breast cancers. *Cancer Research*. **75**(15 Supplement): p. 2948-2948.
45. Swanton, C., et al. (2009). Chromosomal instability determines taxane response. *Proceedings of the National Academy of Sciences of the United States of America*. **106**(21): p. 8671-8676.

46. Tan, D.S., et al. (2013). Implications of BRCA1 and BRCA2 mutations for the efficacy of paclitaxel monotherapy in advanced ovarian cancer. *European Journal of Cancer*. **49**(6): p. 1246-1253.
47. Kriege, M., et al. (2012). The efficacy of taxane chemotherapy for metastatic breast cancer in BRCA1 and BRCA2 mutation carriers. *Cancer*. **118**(4): p. 899-907.
48. Mitchell, G. (2014). 27IN Considerations for the oncologist when treating cancer patients with germline mutations. *Annals of Oncology*. **25**(suppl 4): p. iv11-iv11.
49. Higby, D.J., et al. (1974). Diaminodichloroplatinum: a phase I study showing responses in testicular and other tumors. *Cancer*. **33**(5): p. 1219-1215.
50. Wozniak, A.J., et al. (1998). Randomized trial comparing cisplatin with cisplatin plus vinorelbine in the treatment of advanced non-small-cell lung cancer: a Southwest Oncology Group study. *Journal of Clinical Oncology*. **16**(7): p. 2459-2465.
51. F.D.A. (2007) *Oncology Tools. Listing of approved oncology drugs with approved indications*. [cited 2017 06/09]; Available from: https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm279174.htm.
52. Kelland, L. (2007). The resurgence of platinum-based cancer chemotherapy. *Nature Reviews: Cancer*. **7**(8): p. 573-584.
53. Seiwert, T.Y., J.K. Salama, and E.E. Vokes. (2007). The chemoradiation paradigm in head and neck cancer. *Nature Clinical Practice: Oncology*. **4**(3): p. 156-171.
54. Kelland, L.R. and N.P. Farrell. (2000). *Platinum-based drugs in cancer therapy*. New York: Springer Science & Business Media.
55. Ko, A. (2008). *Everyone's guide to cancer therapy: How cancer is diagnosed, treated, and managed day to day*. 5th ed. Kansas: Andrews McMeel Publishing.
56. Boyiadzis, M., et al. (2014). *Hematology-oncology therapy*. 2nd ed. USA: McGraw-Hill Education.
57. Ruzzo, A., et al. (2007). Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFOX-4 chemotherapy. *Journal of Clinical Oncology*. **25**(10): p. 1247-1254.
58. Lecomte, T., et al. (2006). Glutathione S-transferase P1 polymorphism (Ile105Val) predicts cumulative neuropathy in patients receiving oxaliplatin-based chemotherapy. *Clinical Cancer Research*. **12**(10): p. 3050-3056.
59. Nagashima, F., et al. (2006). Polymorphism in sodium-channel alpha 1-subunit (SCN1A) predicts response, TTP, survival, and toxicity in patients with metastatic colorectal cancer treated with 5-FU/oxaliplatin, in *ASCO Annual Meeting Proceedings.*
60. Corrigan, A., et al. (2014). Pharmacogenetics of pemetrexed combination therapy in lung cancer: pathway analysis reveals novel toxicity associations. *Pharmacogenomics Journal*. **14**(5): p. 411-417.
61. Corrigan, A., et al. (2016). *AQP8 and C7orf57 predict platinum toxicity*, in *In Submission*.
62. Creutzig, U., et al. (2003). [Analysis of causes of death during intensive chemotherapy according to treatment protocol AML-BFM 93]. *Klinische Padiatrie*. **215**(3): p. 151-158.
63. Yamanaka, T., et al. (2007). Predictive value of chemotherapy-induced neutropenia for the efficacy of oral fluoropyrimidine S-1 in advanced gastric carcinoma. *British Journal of Cancer*. **97**(1): p. 37-42.
64. Frei, E., 3rd and G.P. Canellos. (1980). Dose: a critical factor in cancer chemotherapy. *American Journal of Medicine*. **69**(4): p. 585-594.
65. Crawford, J., D.C. Dale, and G.H. Lyman. (2004). Chemotherapy-induced neutropenia: risks, consequences, and new directions for its management. *Cancer*. **100**(2): p. 228-237.
66. Saarto, T., et al. (1997). Haematological toxicity: a marker of adjuvant chemotherapy efficacy in stage II and III breast cancer. *British Journal of Cancer*. **75**(2): p. 301-305.
67. Poikonen, P., et al. (1999). Leucocyte nadir as a marker for chemotherapy efficacy in node-positive breast cancer treated with adjuvant CMF. *British Journal of Cancer*. **80**(11): p. 1763.
68. Mayers, C., T. Panzarella, and I.F. Tannock. (2001). Analysis of the prognostic effects of inclusion in a clinical trial and of myelosuppression on survival after adjuvant chemotherapy for breast carcinoma. *Cancer*. **91**(12): p. 2246-2257.
69. Cameron, D.A., et al. (2003). Moderate neutropenia with adjuvant CMF confers improved survival in early breast cancer. *British Journal of Cancer*. **89**(10): p. 1837-1842.
70. Di Maio, M., et al. (2005). Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. *Lancet Oncology*. **6**(9): p. 669-677.

71. Nelson, M.R., et al. (2016). The genetics of drug efficacy: opportunities and challenges. *Nature Reviews: Genetics*. **17**(4): p. 197-206.
72. Zaïr, Z.M. and D.R. Singer. (2016). Efflux transporter variants as predictors of drug toxicity in lung cancer patients: systematic review and meta-analysis. *Pharmacogenomics*. (0).
73. Kiyotani, K., et al. (2012). A genome-wide association study identifies four genetic markers for hematological toxicities in cancer patients receiving gemcitabine therapy. *Pharmacogenetics and Genomics*. **22**(4): p. 229-235.
74. Diggle, P., et al. (2002). *Analysis of longitudinal data*. 2nd ed. Oxford: Oxford University Press.
75. Laird, N.M. and J.H. Ware. (1982). Random-effects models for longitudinal data. *Biometrics*. **38**(4): p. 963-974.
76. Sibbald, B. and M. Roland. (1998). Understanding controlled trials. Why are randomised controlled trials important? *The BMJ*. **316**(7126): p. 201.
77. Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*. **25**(24): p. 4334-4344.
78. Roberts, C. and D.J. Torgerson. (1999). Understanding controlled trials: baseline imbalance in randomised controlled trials. *The BMJ*. **319**(7203): p. 185.
79. Senn, S.S. (2008). *Statistical issues in drug development*. 2nd ed. Statistics in Practice. Vol. 69. Chichester: John Wiley & Sons.
80. Yu, L.-M., et al. (2010). Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. *Trials*. **11**(59).
81. Kernan, W.N., et al. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*. **52**(1): p. 19-26.
82. Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics*. **35**(2): p. 503-512.
83. Kahan, B.C. and T.P. Morris. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*. **31**(4): p. 328-340.
84. Kahan, B.C. and T.P. Morris. (2013). Adjusting for multiple prognostic factors in the analysis of randomised trials. *BMC Medical Research Methodology*. **13**(1): p. 99.
85. (2013). *Guideline on adjustment for baseline covariates (EMA/295050/2013)*. European Medicines Agency.
86. Trowman, R., et al. (2007). The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. *Journal of Clinical Epidemiology*. **60**(12): p. 1229-1233.
87. Egbewale, B.E., M. Lewis, and J. Sim. (2014). Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Medical Research Methodology*. **14**: p. 49.
88. Riley, R.D., et al. (2013). Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine*. **10**(2): p. e1001380.
89. Abo-Zaid, G., W. Sauerbrei, and R.D. Riley. (2012). Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Medical Research Methodology*. **12**: p. 56.
90. Royston, P., et al. (2009). Prognosis and prognostic research: Developing a prognostic model. *The BMJ*. **338**: p. b604.
91. Cianfrocca, M. and L.J. Goldstein. (2004). Prognostic and predictive factors in early-stage breast cancer. *Oncologist*. **9**(6): p. 606-616.
92. (2002). *The importance of pharmacovigilance: safety monitoring of medicinal products*. World Health Organization.
93. Mangoni, A.A. and S.H. Jackson. (2004). Age-related changes in pharmacokinetics and pharmacodynamics: basic principles and practical applications. *British Journal of Clinical Pharmacology*. **57**(1): p. 6-14.
94. Hammerlein, A., H. Derendorf, and D.T. Lowenthal. (1998). Pharmacokinetic and pharmacodynamic changes in the elderly. Clinical implications. *Clinical Pharmacokinetics*. **35**(1): p. 49-64.
95. Schiller, J.H., et al. (2002). Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *New England Journal of Medicine*. **346**(2): p. 92-98.
96. Kelly, K., et al. (2001). Randomized phase III trial of paclitaxel plus carboplatin versus vinorelbine plus cisplatin in the treatment of patients with advanced non--small-cell lung

cancer: a Southwest Oncology Group trial. *Journal of Clinical Oncology*. **19**(13): p. 3210-3218.
97. Neijt, J.P., et al. (2000). Exploratory phase III study of paclitaxel and cisplatin versus paclitaxel and carboplatin in advanced ovarian cancer. *Journal of Clinical Oncology*. **18**(17): p. 3084-3092.
98. Du Bois, A., et al. (2003). A randomized clinical trial of cisplatin/paclitaxel versus carboplatin/paclitaxel as first-line treatment of ovarian cancer. *Journal of the National Cancer Institute*. **95**(17): p. 1320-1329.
99. Ozols, R.F., et al. (2003). Phase III trial of carboplatin and paclitaxel compared with cisplatin and paclitaxel in patients with optimally resected stage III ovarian cancer: a gynecologic oncology group study. *Journal of Clinical Oncology*. **21**(17): p. 3194-3200.
100. Corrigan, A. (2016). *An investigation of the pharmacogenetic basis of toxicity to platinum chemotherapy agents*. Doctoral thesis. King's College London
101. Michel, P., et al. (2004). Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. *The BMJ*. **328**(7433): p. 199.
102. Christiaans-Dingelhoff, I., et al. (2011). To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports? *BMC Health Services Research*. **11**(1): p. 1.
103. Buccheri, G., D. Ferrigno, and M. Tamburini. (1996). Karnofsky and ECOG performance status scoring in lung cancer: a prospective, longitudinal study of 536 patients from a single institution. *European Journal of Cancer*. **32A**(7): p. 1135-1141.
104. Oken, M.M., et al. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology*. **5**(6): p. 649-655.
105. (2010). *Common terminology criteria for adverse events (CTCAE) version 4.0*. US Department of Health and Human Services.
106. Capewell, S. and M.F. Sudlow. (1990). Performance and prognosis in patients with lung cancer. The Edinburgh Lung Cancer Group. *Thorax*. **45**(12): p. 951-956.
107. Tummarello, D., et al. (1995). Symptomatic, stage IV, non-small-cell lung cancer (NSCLC): response, toxicity, performance status change and symptom relief in patients treated with cisplatin, vinblastine and mitomycin-C. *Cancer Chemotherapy and Pharmacology*. **35**(3): p. 249-253.
108. Orr, S. and J. Aisner. (1986). Performance status assessment among oncology patients: a review. *Cancer Treatment Reports*. **70**(12): p. 1423-1429.
109. (1997). Clinical practice guidelines for the treatment of unresectable non-small-cell lung cancer. Adopted on May 16, 1997 by the American Society of Clinical Oncology. *Journal of Clinical Oncology*. **15**(8): p. 2996-3018.
110. Bunn, P.A., Jr. (2002). Chemotherapy for advanced non-small-cell lung cancer: who, what, when, why? *Journal of Clinical Oncology*. **20**(18 Suppl): p. 23S-33S.
111. Gridelli, C., et al. (2004). Treatment of advanced non-small-cell lung cancer patients with ECOG performance status 2: results of an European Experts Panel. *Annals of Oncology*. **15**(3): p. 419-426.
112. Hotta, K., et al. (2004). Role of adjuvant chemotherapy in patients with resected non-small-cell lung cancer: reappraisal with a meta-analysis of randomized controlled trials. *Journal of Clinical Oncology*. **22**(19): p. 3860-3867.
113. Ardizzoni, A., et al. (2007). Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis. *Journal of the National Cancer Institute*. **99**(11): p. 847-857.
114. Jiang, J., et al. (2007). A meta-analysis of randomized controlled trials comparing carboplatin-based to cisplatin-based chemotherapy in advanced non-small cell lung cancer. *Lung Cancer*. **57**(3): p. 348-358.
115. Allen, J., et al. (2013). Integrating and extending cohort studies: lessons from the eXtending Treatments, Education and Networks in Depression (xTEND) study. *BMC Medical Research Methodology*. **13**: p. 122.
116. Abebe, H.T., et al. (2015). Bayesian design for dichotomous repeated measurements with autocorrelation. *Statistical Methods in Medical Research*. **24**(5): p. 594-611.
117. Salzberg, M., et al. (2005). Current concepts of treatment strategies in advanced or recurrent ovarian cancer. *Oncology*. **68**(4-6): p. 293-298.
118. Mcguire, W.P. and R.F. Ozols. (1998). Chemotherapy of advanced ovarian cancer. *Seminars in Oncology*. **25**(3): p. 340-348.

119. Markman, M., et al. (2001). Phase III trial of standard-dose intravenous cisplatin plus paclitaxel versus moderately high-dose carboplatin followed by intravenous paclitaxel and intraperitoneal cisplatin in small-volume stage III ovarian carcinoma: an intergroup study of the Gynecologic Oncology Group, Southwestern Oncology Group, and Eastern Cooperative Oncology Group. *Journal of Clinical Oncology*. **19**(4): p. 1001-1007.
120. Katsumata, N., et al. (2009). Dose-dense paclitaxel once a week in combination with carboplatin every 3 weeks for advanced ovarian cancer: a phase 3, open-label, randomised controlled trial. *Lancet*. **374**(9698): p. 1331-1338.
121. Brade, A., et al. (2016). Phase II study of concurrent pemetrexed, cisplatin, and radiation therapy for stage IIIA/B unresectable non-small cell lung cancer. *Clinical Lung Cancer*. **17**(2): p. 133-141.
122. Shukuya, T., et al. (2015). Nedaplatin plus docetaxel versus cisplatin plus docetaxel for advanced or relapsed squamous cell carcinoma of the lung (WJOG5208L): a randomised, open-label, phase 3 trial. *Lancet Oncology*. **16**(16): p. 1630-1638.
123. Hoang, T., et al. (2005). Clinical model to predict survival in chemonaive patients with advanced non-small-cell lung cancer treated with third-generation chemotherapy regimens based on eastern cooperative oncology group data. *Journal of Clinical Oncology*. **23**(1): p. 175-183.
124. Hoang, T., et al. (2012). Prognostic models to predict survival in non-small-cell lung cancer patients treated with first-line paclitaxel and carboplatin with or without bevacizumab. *Journal of Thoracic Oncology*. **7**(9): p. 1361-1368.
125. Omura, G.A., et al. (1991). Long-term follow-up and prognostic factor analysis in advanced ovarian carcinoma: the gynecologic oncology group experience. *Journal of Clinical Oncology*. **9**(7): p. 1138-1150.
126. Holschneider, C.H. and J.S. Berek. (2000). Ovarian cancer: epidemiology, biology, and prognostic factors. *Seminars in Surgical Oncology*. **19**(1): p. 3-10.
127. Tingulstad, S., et al. (2003). Survival and prognostic factors in patients with ovarian cancer. *Obstetrics and Gynecology*. **101**(5 Pt 1): p. 885-891.
128. Tinquaut, F., et al. (2016). Prognostic factors for overall survival in elderly patients with advanced ovarian cancer treated with chemotherapy: results of a pooled analysis of three GINECO phase II trials. *Gynecologic Oncology*. **143**(1): p. 22-26.
129. Ray-Coquard, I., et al. (2009). Lymphopenia as a prognostic factor for overall survival in advanced carcinomas, sarcomas, and lymphomas. *Cancer Research*. **69**(13): p. 5383-5391.
130. Lee, H.J., J.Y. Moon, and S.W. Baek. (2016). Is treatment-emergent toxicity a biomarker of efficacy of apatinib in gastric cancer? *Journal of Clinical Oncology*. p. JCO688663.
131. Nakamura, M., et al. (2016). Phase II study for determining usefulness of FDG-PET as imaging biomarker in regorafenib treatment for metastatic colorectal cancer (JACCRO CC-12), in *ASCO Annual Meeting Proceedings.*
132. Hellman, S., S.A. Rosenberg, and V.T. Devita. (1997). *Cancer: principles & practice of oncology*. Vol. 2. USA: Lippincott-Raven Publishers.
133. Caggiano, V., et al. (2005). Incidence, cost, and mortality of neutropenia hospitalization associated with chemotherapy. *Cancer*. **103**(9): p. 1916-1924.
134. Ricci, M.S. and W.X. Zong. (2006). Chemotherapeutic approaches for targeting cell death pathways. *Oncologist*. **11**(4): p. 342-357.
135. Dean, M., T. Fojo, and S. Bates. (2005). Tumour stem cells and drug resistance. *Nature Reviews: Cancer*. **5**(4): p. 275-284.
136. Shitara, K., et al. (2009). Neutropaenia as a prognostic factor in metastatic colorectal cancer patients undergoing chemotherapy with first-line FOLFOX. *European Journal of Cancer*. **45**(10): p. 1757-1763.
137. Samson, M.K., et al. (1984). Dose-response and dose-survival advantage for high versus low-dose cisplatin combined with vinblastine and bleomycin in disseminated testicular cancer. A Southwest Oncology Group study. *Cancer*. **53**(5): p. 1029-1035.
138. Rankin, E.M., et al. (1992). A randomised study comparing standard dose carboplatin with chlorambucil and carboplatin in advanced ovarian cancer. *British Journal of Cancer*. **65**(2): p. 275-281.
139. Hovgaard, D. and N. Nissen. (1992). A phase I/II study of dose and administration of non-glycosylated bacterially synthesized GM CSF in chemotherapy-induced neutropenia in patients with non-Hodgkin's lymphomas. *Leukemia & Lymphoma*. **7**(3): p. 217-224.

140. Kishida, Y., et al. (2009). Chemotherapy-induced neutropenia as a prognostic factor in advanced non-small-cell lung cancer: results from Japan Multinational Trial Organization LC00-03. *British Journal of Cancer*. **101**(9): p. 1537-1542.
141. Yamanaka, T., et al. (2007). Predictive value of chemotherapy-induced neutropenia for drug efficacy in advanced gastric carcinoma: Analysis of prospective nationwide survey, in *ASCO Annual Meeting Proceedings.*
142. Kawahara, T., et al. (2016). Neutrophil-to-lymphocyte ratio is a prognostic marker in bladder cancer patients after radical cystectomy. *BMC Cancer*. **16**(1): p. 1.
143. Demirtaş, A., et al. (2013). Can neutrophil-lymphocyte ratio and lymph node density be used as prognostic factors in patients undergoing radical cystectomy? *The Scientific World Journal*. **2013**: p. 1-5.
144. Xue, P., et al. (2014). Neutrophil-to-lymphocyte ratio for predicting palliative chemotherapy outcomes in advanced pancreatic cancer patients. *Cancer Medicine*. **3**(2): p. 406-415.
145. Saad, E., et al. (2009). Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Annals of Oncology*. **21**(1): p. 7-12.
146. Zietemann, V.D., T. Schuster, and T.H. Duell. (2011). Post-study therapy as a source of confounding in survival analysis of first-line studies in patients with advanced non-small-cell lung cancer. *Journal of Thoracic Disease*. **3**(2): p. 88.
147. Cox, D.R. (1992). *Regression models and life-tables*, in *Breakthroughs in statistics*. New York: Springer-Verlag, p. 527-541.
148. Kaplan, E.L. and P. Meier. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*. **53**(282): p. 457-481.
149. Goel, M.K., P. Khanna, and J. Kishore. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*. **1**(4): p. 274-278.
150. Bellera, C.A., et al. (2010). Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*. **10**(1): p. 1.
151. Powell, T.M. and M.E. Bagnell. (2012). Your "survival" guide to using time-dependent covariates, in *Proceedings of the SAS Global Forum.* Citeseer.
152. Kleinbaum, D.G. and M. Klein. (2006). *Survival analysis: a self-learning text*. 2nd ed. Statistics for Biology and Health. USA: Springer Science & Business Media.
153. Therneau, T.M. (2014) *Cox models and "type III" tests*. Mayo Clinic:[Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.6232&rep=rep1&type=pdf.
154. S.A.S. (2009). *SAS/STAT® 9.2. User's Guide*. SAS Institute Inc: Cary, NC.
155. Nakata, B., et al. (2006). Moderate neutropenia with S-1 plus low-dose cisplatin may predict a more favourable prognosis in advanced gastric cancer. *Clinical Oncology*. **18**(9): p. 678-683.
156. Mateos, M.V., et al. (2015). Effect of cumulative bortezomib dose on survival in multiple myeloma patients receiving bortezomib-melphalan-prednisone in the phase III VISTA study. *American Journal of Hematology*. **90**(4): p. 314-319.
157. Rambach, L., et al. (2014). Prognostic value of chemotherapy-induced hematological toxicity in metastatic colorectal cancer patients. *World Journal of Gastroenterology*. **20**(6): p. 1565-1573.
158. Morden, N.E., et al. (2012). End-of-life care for Medicare beneficiaries with cancer is highly intensive overall and varies widely. *Health Affairs*. **31**(4): p. 786-796.
159. Carus, A., et al. (2013). Impact of baseline and nadir neutrophil index in non-small cell lung cancer and ovarian cancer patients: Assessment of chemotherapy for resolution of unfavourable neutrophilia. *Journal of Translational Medicine*. **11**(1): p. 189.
160. Chen, Z., et al. (2015). Pretreated baseline neutrophil count and chemotherapy-induced neutropenia may be conveniently available as prognostic biomarkers in advanced gastric cancer. *Internal Medicine Journal*. **45**(8): p. 854-859.
161. Crawford, J., et al. (1991). Reduction by granulocyte colony-stimulating factor of fever and neutropenia induced by chemotherapy in patients with small-cell lung cancer. *New England Journal of Medicine*. **325**(3): p. 164-170.
162. Smith, T.J., et al. (2015). Recommendations for the use of WBC growth factors: American Society of Clinical Oncology clinical practice guideline update. *Journal of Clinical Oncology*. **33**(28): p. 3199-3212.

163. Lyman, G.H., et al. (2003). Risk of febrile neutropenia among patients with intermediate-grade non-Hodgkin's lymphoma receiving CHOP chemotherapy. *Leukemia & Lymphoma*. **44**(12): p. 2069-2076.
164. Hosmer, W., J. Malin, and M. Wong. (2011). Development and validation of a prediction model for the risk of developing febrile neutropenia in the first cycle of chemotherapy among elderly patients with breast, lung, colorectal, and prostate cancer. *Supportive Care in Cancer*. **19**(3): p. 333-341.
165. Lyman, G.H., et al. (2011). Predicting individual risk of neutropenic complications in patients receiving cancer chemotherapy. *Cancer*. **117**(9): p. 1917-1927.
166. Yabroff, K.R., et al. (2008). Cost of care for elderly cancer patients in the United States. *Journal of the National Cancer Institute*. **100**(9): p. 630-641.
167. Hassett, M.J., et al. (2006). Frequency and cost of chemotherapy-related serious adverse effects in a population sample of women with breast cancer. *Journal of the National Cancer Institute*. **98**(16): p. 1108-1117.
168. Brooks, G.A., et al. (2015). A clinical prediction model to assess risk for chemotherapy-related hospitalization in patients initiating palliative chemotherapy. *JAMA Oncology*. **1**(4): p. 441-447.
169. Brooks, G.A., et al. (2014). Identification of potentially avoidable hospitalizations in patients with GI cancer. *Journal of Clinical Oncology*. **32**(6): p. 496-503.
170. Hassett, M.J., et al. (2011). Chemotherapy-related hospitalization among community cancer center patients. *Oncologist*. **16**(3): p. 378-387.
171. Hurria, A., et al. (2011). Predicting chemotherapy toxicity in older adults with cancer: a prospective multicenter study. *Journal of Clinical Oncology*. **29**(25): p. 3457-3465.
172. Extermann, M., et al. (2012). Predicting the risk of chemotherapy toxicity in older patients: the Chemotherapy Risk Assessment Scale for High-Age Patients (CRASH) score. *Cancer*. **118**(13): p. 3377-3386.
173. Hyman, D.M., et al. (2014). Nomogram to predict cycle-one serious drug-related toxicity in phase I oncology trials. *Journal of Clinical Oncology*. **32**(6): p. 519-526.
174. Osterlind, K. and P.K. Andersen. (1986). Prognostic factors in small cell lung cancer: multivariate model based on 778 patients treated with chemotherapy with or without irradiation. *Cancer Research*. **46**(8): p. 4189-4194.
175. Chen, X., et al. (2014). Multivariate analysis of prognostic factors in the eldly patients with small cell lung cancer: a study of 160 patients. *Zhongguo Fei Ai Za Zhi/Chinese Journal of Lung Cancer*. **17**(1): p. 15-23.
176. Forrest, L.M., et al. (2004). Comparison of an inflammation-based prognostic score (GPS) with performance status (ECOG) in patients receiving platinum-based chemotherapy for inoperable non-small-cell lung cancer. *British Journal of Cancer*. **90**(9): p. 1704-1706.
177. Carey, M.S., et al. (2008). The prognostic effects of performance status and quality of life scores on progression-free survival and overall survival in advanced ovarian cancer. *Gynecologic Oncology*. **108**(1): p. 100-105.
178. Hofman, M., et al. (2007). Cancer-related fatigue: the scale of the problem. *Oncologist*. **12 Suppl 1**(Supplement 1): p. 4-10.
179. Lin, J. and G.C. Curhan. (2008). Kidney function decline and physical function in women. *Nephrology, Dialysis, Transplantation*. **23**(9): p. 2827-2833.
180. Johansen, K.L., et al. (2003). Longitudinal study of nutritional status, body composition, and physical function in hemodialysis patients. *American Journal of Clinical Nutrition*. **77**(4): p. 842-846.
181. Kurella, M., et al. (2004). Physical and sexual function in women with chronic kidney disease. *American Journal of Kidney Diseases*. **43**(5): p. 868-876.
182. Odden, M.C., M.A. Whooley, and M.G. Shlipak. (2004). Association of chronic kidney disease and anemia with physical capacity: the heart and soul study. *Journal of the American Society of Nephrology*. **15**(11): p. 2908-2915.
183. Odden, M.C., et al. (2006). Cystatin C and measures of physical function in elderly adults: the Health, Aging, and Body Composition (HABC) Study. *American Journal of Epidemiology*. **164**(12): p. 1180-1189.
184. Mcmanus, D., et al. (2007). Association of cystatin C with poor exercise capacity and heart rate recovery: data from the heart and soul study. *American Journal of Kidney Diseases*. **49**(3): p. 365-372.
185. Hogberg, T., J. Carstensen, and E. Simonsen. (1993). Treatment results and prognostic factors in a population-based study of epithelial ovarian cancer. *Gynecologic Oncology*. **48**(1): p. 38-49.

186. Levi, F., et al. (1993). Epidemiologic pathology of ovarian cancer from the Vaud Cancer Registry, Switzerland. *Annals of Oncology*. **4**(4): p. 289-294.
187. Malkasian, G.D., Jr., et al. (1984). Prognostic significance of histologic classification and grading of epithelial malignancies of the ovary. *American Journal of Obstetrics and Gynecology*. **149**(3): p. 274-284.
188. Janssen-Heijnen, M.L., et al. (2004). Effect of comorbidity on the treatment and prognosis of elderly patients with non-small cell lung cancer. *Thorax*. **59**(7): p. 602-607.
189. Asmis, T.R., et al. (2008). Age and comorbidity as independent prognostic factors in the treatment of non–small-cell lung cancer: a review of National Cancer Institute of Canada Clinical Trials Group trials. *Journal of Clinical Oncology*. **26**(1): p. 54-59.
190. Gridelli, C., et al. (2007). Lung cancer in the elderly. *Journal of Clinical Oncology*. **25**(14): p. 1898-1907.
191. Moyer, V.A. (2014). Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*. **160**(5): p. 330-338.
192. Alberg, A.J., et al. (2013). Epidemiology of lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*. **143**(5 Suppl): p. e1S-e29S.
193. Lee, J.-Y., et al. (2013). Diabetes mellitus as an independent risk factor for lung cancer: a meta-analysis of observational studies. *European Journal of Cancer*. **49**(10): p. 2411-2423.
194. Ouellette, D., et al. (1998). Lung cancer in women compared with men: stage, treatment, and survival. *Annals of Thoracic Surgery*. **66**(4): p. 1140-1143; discussion 1143-1144.
195. Wisnivesky, J.P. and E.A. Halm. (2007). Sex differences in lung cancer survival: do tumors behave differently in elderly women? *Journal of Clinical Oncology*. **25**(13): p. 1705-1712.
196. Wheatley-Price, P., et al. (2010). The influence of sex on efficacy, adverse events, quality of life, and delivery of treatment in National Cancer Institute of Canada Clinical Trials Group non-small cell lung cancer chemotherapy trials. *Journal of Thoracic Oncology*. **5**(5): p. 640-648.
197. Radzikowska, E., P. Glaz, and K. Roszkowski. (2002). Lung cancer in women: age, smoking, histology, performance status, stage, initial treatment and survival. Population-based study of 20 561 cases. *Annals of Oncology*. **13**(7): p. 1087-1093.
198. Tammemagi, C.M., et al. (2004). In lung cancer patients, age, race-ethnicity, gender and smoking predict adverse comorbidity, which in turn predicts treatment and survival. *Journal of Clinical Epidemiology*. **57**(6): p. 597-609.
199. Stein, B.N., et al. (1995). Age and sex are independent predictors of 5-fluorouracil toxicity. Analysis of a large scale phase III trial. *Cancer*. **75**(1): p. 11-17.
200. Zalcberg, J., et al. (1998). Haematological and non-haematological toxicity after 5-fluorouracil and leucovorin in patients with advanced colorectal cancer is significantly associated with gender, increasing age and cycle number. Tomudex International Study Group. *European Journal of Cancer*. **34**(12): p. 1871-1875.
201. Pocock, S.J., N.L. Geller, and A.A. Tsiatis. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*. **43**(3): p. 487-498.
202. Teixeira-Pinto, A., et al. (2009). Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric Annals*. **39**(7): p. 729-735.
203. Dabrowska, D.M. and K.A. Doksum. (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*. **83**(403): p. 744-749.
204. Prentice, R.L. and J. Cai. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*. **79**(3): p. 495-512.
205. Oakes, D. (1994). Multivariate survival distributions. *Journal of Nonparametric Statistics*. **3**(3-4): p. 343-354.
206. Huberty, C.J. and J.D. Morris. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*. **105**(2): p. 302.
207. Robinson, L.D. and N.P. Jewell. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*. p. 227-240.
208. Pirinen, M., P. Donnelly, and C.C. Spencer. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*. **44**(8): p. 848-851.

209. Nuijten, M., et al. (2010). Effectiveness of bevacizumab-and pemetrexed-cisplatin treatment for patients with advanced non-squamous non-small cell lung cancer. *Lung Cancer*. **69**: p. S4-S10.
210. Group, I.a.L.C.T.C. (2004). Cisplatin-based adjuvant chemotherapy in patients with completely resected non–small-cell lung cancer. *New England Journal of Medicine*. **2004**(350): p. 351-360.
211. Schmid-Bindert, G., et al. (2015). A randomized Phase 2 study of pemetrexed in combination with cisplatin or carboplatin as adjuvant chemotherapy in patients with completely resected stage IB or II Non-Small-Cell Lung Cancer. *Lung Cancer*. **90**(3): p. 397-404.
212. Petrillo, M., et al. (2016). Cytoreductive surgery plus platinum-based hyperthermic intraperitoneal chemotherapy in epithelial ovarian cancer: a promising integrated approach to improve locoregional control. *Oncologist*. **21**(5): p. 532-534.
213. Zivanovic, O., et al. (2015). HIPEC ROC I: A phase i study of cisplatin administered as hyperthermic intraoperative intraperitoneal chemoperfusion followed by postoperative intravenous platinum-based chemotherapy in patients with platinum-sensitive recurrent epithelial ovarian cancer. *International Journal of Cancer*. **136**(3): p. 699-708.
214. Avan, A., et al. (2015). Platinum-induced neurotoxicity and preventive strategies: past, present, and future. *Oncologist*. **20**(4): p. 411-432.
215. Kalikaki, A., et al. (2015). ERCC1 SNPs as potential predictive biomarkers in non-small cell lung cancer patients treated with platinum-based chemotherapy. *Cancer Investigation*. **33**(4): p. 107-113.
216. Wells, Q.S., J.T. Delaney, and D.M. Roden. (2012). Genetic determinants of response to cardiovascular drugs. *Current Opinion in Cardiology*. **27**(3): p. 253-261.
217. Crawford, D.C., M.D. Ritchie, and M.J. Rieder. (2007). Identifying the genotype behind the phenotype: a role model found in VKORC1 and its association with warfarin dosing. *Pharmacogenomics*. **8**(5): p. 487-496.
218. Tan, X.-L., et al. (2011). Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clinical Cancer Research*. **17**(17): p. 5801-5811.
219. Hieke, S., et al. (2016). Integrating multiple molecular sources into a clinical risk prediction signature by extracting complementary information. *BMC Bioinformatics*. **17**(1): p. 327.
220. Gondos, A., et al. (2006). Calculating age-adjusted cancer survival estimates when age-specific data are sparse: an empirical evaluation of various methods. *British Journal of Cancer*. **94**(3): p. 450-454.
221. Brenner, H. and T. Hakulinen. (2005). Age adjustment of cancer survival rates: methods, point estimates and standard errors. *British Journal of Cancer*. **93**(3): p. 372-375.
222. Tas, F., et al. (2013). Age is a prognostic factor affecting survival in lung cancer patients. *Oncology Letters*. **6**(5): p. 1507-1513.
223. Winter, W.E., 3rd, et al. (2007). Prognostic factors for stage III epithelial ovarian cancer: a gynecologic oncology group study. *Journal of Clinical Oncology*. **25**(24): p. 3621-3627.
224. Landrum, L.M., et al. (2013). Prognostic factors for stage III epithelial ovarian cancer treated with intraperitoneal chemotherapy: a gynecologic oncology group study. *Gynecologic Oncology*. **130**(1): p. 12-18.
225. Visbal, A.L., et al. (2004). Gender differences in non–small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. *Annals of Thoracic Surgery*. **78**(1): p. 209-215.
226. Cagle, P.T. and L.R. Chirieac. (2012). Advances in treatment of lung cancer with targeted therapy. *Archives of Pathology and Laboratory Medicine*. **136**(5): p. 504-509.
227. Leger, K.J., et al. (2016). Clinical and genetic determinants of cardiomyopathy risk among hematopoietic cell transplantation survivors. *Biology of Blood and Marrow Transplantation*. **22**(6): p. 1094-1101.
228. Sombekke, M.H., et al. (2010). Analysis of multiple candidate genes in association with phenotypes of multiple sclerosis. *Multiple Sclerosis*. **16**(6): p. 652-659.
229. Bovelstad, H.M., S. Nygard, and O. Borgan. (2009). Survival prediction from clinico-genomic models--a comparative study. *BMC Bioinformatics*. **10**(1): p. 413.
230. Vazquez, A.I., et al. (2016). Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. *Genetics*. **203**(3): p. 1425-1438.
231. Neuhaus, J.M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association*. **93**(443): p. 1124-1129.

232. Bøvelstad, H.M., S. Nygård, and Ø. Borgan. (2009). Survival prediction from clinico-genomic models-a comparative study. *BMC Bioinformatics*. **10**(1): p. 413.
233. Hsieh, F.Y. and P.W. Lavori. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*. **21**(6): p. 552-560.
234. Schoenfeld, D.A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*. **39**(2): p. 499-503.
235. Purcell, S., S.S. Cherny, and P.C. Sham. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. **19**(1): p. 149-150.
236. Bacchetti, P. and J.M. Leung. (2002). Sample size calculations in clinical research. *Anesthesiology*. **97**(4): p. 1028-1029; author reply 1029-1032.
237. Kalbfleisch, J.D. and R.L. Prentice. (2011). *The statistical analysis of failure time data*. 2nd ed. Vol. 360. Hoboken, NJ: John Wiley & Sons.
238. Schemper, M. and T.L. Smith. (1996). A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials*. **17**(4): p. 343-346.
239. Cook, J.P. and A.P. Morris. (2016). Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal of Human Genetics*. **24**(8): p. 1175-1180.
240. Klein, J.P. and M.L. Moeschberger. (2005). *Survival analysis: techniques for censored and truncated data*. 2nd ed. Statistics for Biology and Health. USA: Springer Science & Business Media.
241. Kleinbaum, D.G. and M. Klein. (2010). *Survival analysis*. 3rd ed. Statistics for Biology and Health. USA: Springer.
242. Aschard, H., et al. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *American Journal of Human Genetics*. **96**(2): p. 329-339.
243. Kahan, B.C., et al. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. **15**(1): p. 139.
244. Pocock, S.J., et al. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine*. **21**(19): p. 2917-2930.
245. Hsieh, F.Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*. **8**(7): p. 795-802.
246. Hsieh, F.Y., D.A. Bloch, and M.D. Larsen. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*. **17**(14): p. 1623-1634.
247. Hsieh, F.Y., et al. (2003). An overview of variance inflation factors for sample-size calculation. *Evaluation and the Health Professions*. **26**(3): p. 239-257.
248. Wong, K.M., T.J. Hudson, and J.D. Mcpherson. (2011). Unraveling the genetics of cancer: genome sequencing and beyond. *Annual Review of Genomics and Human Genetics*. **12**: p. 407-430.
249. Kong, B., et al. (2014). p53 is required for cisplatin-induced processing of the mitochondrial fusion protein L-Opa1 that is mediated by the mitochondrial metallopeptidase Oma1 in gynecologic cancers. *Journal of Biological Chemistry*. **289**(39): p. 27134-27145.
250. Wang, F., et al. (2014). Loss of TACSTD2 contributed to squamous cell carcinoma progression through attenuating TAp63-dependent apoptosis. *Cell Death & Disease*. **5**(3): p. e1133.
251. Gutman, S.I., et al. (2013). *Progression-free survival: what does it mean for psychological well-being or quality of life?* Methods Research Reports. Rockville, MD: Agency for Healthcare Research and Quality.
252. Gratten, J. and P.M. Visscher. (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine*. **8**(1): p. 78.
253. Begg, C.B. (2013). Justifying the choice of endpoints for clinical trials. *Journal of the National Cancer Institute*. **105**(21): p. 1594-1595.
254. Tan, X.L., et al. (2011). Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clinical Cancer Research*. **17**(17): p. 5801-5811.
255. Lose, F., et al. (2010). Vascular endothelial growth factor gene polymorphisms and ovarian cancer survival. *Gynecologic Oncology*. **119**(3): p. 479-483.

256. Xun, W.W., et al. (2011). Single-nucleotide polymorphisms (5p15.33, 15q25.1, 6p22.1, 6q27 and 7p15.3) and lung cancer survival in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Mutagenesis*. **26**(5): p. 657-666.
257. Wu, X., et al. (2013). Genome-wide association study of genetic predictors of overall survival for non–small cell lung cancer in never smokers. *Cancer Research*. **73**(13): p. 4028-4038.
258. Ratner, E., et al. (2010). A KRAS-variant in ovarian cancer acts as a genetic marker of cancer risk. *Cancer Research*. **70**(16): p. 6509-6515.
259. French, J.D., et al. (2016). Germline polymorphisms in an enhancer of PSIP1 are associated with progression-free survival in epithelial ovarian cancer. *Oncotarget*. **7**(6): p. 6353-6368.
260. Huang, R.S., et al. (2011). Platinum sensitivity-related germline polymorphism discovered via a cell-based approach and analysis of its association with outcome in ovarian cancer patients. *Clinical Cancer Research*. **17**(16): p. 5490-5500.
261. Kerns, S.L., et al. (2016). Meta-analysis of Genome Wide Association Studies Identifies Genetic Markers of Late Toxicity Following Radiotherapy for Prostate Cancer. *EBioMedicine*. **10**: p. 150-163.
262. Rosmarin, D., et al. (2015). A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at DPYD and a putative role for ENOSF1 rather than TYMS. *Gut*. **64**(1): p. 111-120.
263. Dorling, L., et al. (2016). The relationship between common genetic markers of breast cancer risk and chemotherapy-induced toxicity: a case-control study. *PloS One*. **11**(7): p. e0158984.
264. Gréen, H., et al. (2016). Using whole-exome sequencing to identify genetic markers for carboplatin and gemcitabine-induced toxicities. *Clinical Cancer Research*. **22**(2): p. 366-373.
265. Cao, S., et al. (2016). Genome-wide association study of myelosuppression in non-small-cell lung cancer patients with platinum-based chemotherapy. *Pharmacogenomics Journal*. **16**(1): p. 41-46.
266. O'reilly, P.F., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS One*. **7**(5): p. e34861.
267. Kim, S., K.A. Sohn, and E.P. Xing. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*. **25**(12): p. i204-212.
268. Medland, S.E. and M.C. Neale. (2010). An integrated phenomic approach to multivariate allelic association. *European Journal of Human Genetics*. **18**(2): p. 233-239.
269. Berg, J.J. and G. Coop. (2014). A population genetic signal of polygenic adaptation. *PLoS Genetics*. **10**(8): p. e1004412.
270. Galesloot, T.E., et al. (2014). A comparison of multivariate genome-wide association methods. *PloS One*. **9**(4): p. e95923.
271. Willer, C.J., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*. **40**(2): p. 161-169.
272. Yang, Q., et al. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*. **34**(5): p. 444-454.
273. Amrhein, V., F. Korner-Nievergelt, and T. Roth. (2017). The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ Preprints*. p. 5:e2921v2922, https://doi.org/2910.7287/peerj.preprints.2921v2922.
274. Pritschet, L., D. Powell, and Z. Horne. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*. **27**(7): p. 1036-1042.
275. Fisher, R.A. (1956). *Statistical methods and scientific inference*. Oxford: Hafner.
276. Fisher, R.A. (1958). *Statistical methods for research workers*. 13th ed. Edinburgh: Oliver and Boyd Ltd.
277. Drummond, G. (2015). Most of the time, P is an unreliable marker, so we need no exact cut-off. *British Journal of Anaesthesia*. **116**(6): p. 894-894.
278. Higgs, M.D. (2013). Do we really need the s-word. *American Scientist*. **101**(1): p. 6-9.
279. Van Helden, J. (2016). Confidence intervals are no salvation from the alleged fickleness of the P value. *Nature Methods*. **13**(8): p. 605-606.
280. De Castria, T.B., et al. (2013). Cisplatin versus carboplatin in combination with third-generation drugs for advanced non-small cell lung cancer. *Cochrane Database Syst Rev*. (8): p. CD009256.

281. Rosell, R., et al. (2002). Phase III randomised trial comparing paclitaxel/carboplatin with paclitaxel/cisplatin in patients with advanced non-small-cell lung cancer: a cooperative multinational trial. *Annals of Oncology*. **13**(10): p. 1539-1549.
282. Arnold, B.F., et al. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*. **11**(1): p. 94.
283. Guo, Y., et al. (2013). Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology*. **13**(1): p. 100.
284. Meldrum, M.L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America*. **14**(4): p. 745-760.
285. Treasure, T. and K.D. Macrae. (1998). Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *The BMJ*. **317**(7155): p. 362.
286. Jadad, A.R., M. Enkin, and A.R. Jadad. (2007). *Randomized controlled trials: questions, answers, and musings*. 2nd ed. Malden, MA: Blackwell Publishing.
287. Jensen, D. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics*. p. 813-825.
288. Ryan, T.P. (2013). *Sample size determination and power*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.
289. Broglio, K.R., J.T. Connor, and S.M. Berry. (2014). Not too big, not too small: A Goldilocks approach to sample size selection. *Journal of Biopharmaceutical Statistics*. **24**(3): p. 685-705.
290. Molenberghs, G., et al. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. **5**(3): p. 445-464.
291. Vickers, A.J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology*. **1**(1): p. 6.
292. Overall, J.E. and R.R. Starbuck. (1979). Sample size estimation for randomized pre-post designs. *Journal of Psychiatric Research*. **15**(1): p. 51-55.
293. Overall, J.E. (1987). Estimating sample size for longitudinal studies of age-related cognitive decline. *Journal of Gerontology*. **42**(2): p. 137-141.
294. Fleiss, J.L. (2011). *Design and analysis of clinical experiments*. Vol. 73. John Wiley & Sons.
295. Ahn, C., M. Heo, and S. Zhang. (2014). *Sample size calculations for clustered and longitudinal outcomes in clinical research*. Boca Raton, FL: CRC Press.
296. Allison, P.D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*. p. 93-114.
297. Vickers, A.J. and D.G. Altman. (2001). Statistics notes: Analysing controlled trials with baseline and follow up measurements. *The BMJ*. **323**(7321): p. 1123-1124.
298. Bryan, M. and P.J. Heagerty. (2014). Direct regression models for longitudinal rates of change. *Statistics in Medicine*. **33**(12): p. 2115-2136.
299. Zhang, S. and C. Ahn. (2011). How many measurements for time-averaged differences in repeated measurement studies? *Contemporary Clinical Trials*. **32**(3): p. 412-417.
300. Overall, J.E. and S.R. Doyle. (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials*. **15**(2): p. 100-123.
301. Overall, J.E. (1996). How many repeated measurements are useful? *Journal of Clinical Psychology*. **52**(3): p. 243-252.
302. Fuguitt, G.V. and S. Lieberson. (1973). Correlation of ratios or difference scores having common terms. *Sociological Methodology*. **5**: p. 128-144.
303. Frison, L. and S.J. Pocock. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine*. **11**(13): p. 1685-1704.
304. Overall, J.E., G. Shobaki, and C.B. Anderson. (1998). Comparative evaluation of two models for estimating sample sizes for tests on trends across repeated measurements. *Controlled Clinical Trials*. **19**(2): p. 188-197.
305. Ahn, C., J.E. Overall, and S. Tonidandel. (2001). Sample size and power calculations in repeated measurement analysis. *Computer Methods and Programs in Biomedicine*. **64**(2): p. 121-124.
306. Yim, L.S. (2012). Baseline adjustment for statistical efficiency on clinical controlled trial. *Journal of Health and Translational Medicine*. **12**(1).

307. Salinsky, M.C., et al. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12-16-week period. *Journal of the International Neuropsychological Society*. **7**(5): p. 597-605.
308. Portney, L. and M. Watkins. (2000). *Foundations of clinical research: applications to practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall Health.
309. O'brien, R.G. and M.K. Kaiser. (1985). MANOVA method for analyzing repeated measures designs: an extensive primer. *Psychological Bulletin*. **97**(2): p. 316-333.
310. Pan, W. (2001). Sample size and power calculations with correlated binary data. *Controlled Clinical Trials*. **22**(3): p. 211-227.
311. Liu, G. and K.Y. Liang. (1997). Sample size calculations for studies with correlated observations. *Biometrics*. **53**(3): p. 937-947.
312. Lu, K., D.V. Mehrotra, and G. Liu. (2009). Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine*. **28**(4): p. 679-699.
313. Hedeker, D., R.D. Gibbons, and C. Waternaux. (1999). Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*. **24**(1): p. 70-93.
314. Kapur, K., et al. (2014). Sample size determination for longitudinal designs with binary response. *Statistics in Medicine*. **33**(22): p. 3781-3800.
315. Morgan, T.M. and L.D. Case. (2013). Conservative sample size determination for repeated measures analysis of covariance. *Annals of Biometrics and Biostatistics*. **1**(1).
316. Nayak, B.K. (2010). Understanding the relevance of sample size calculation. *Indian Journal of Ophthalmology*. **58**(6): p. 469-470.
317. Machin, D., et al. (2011). *Sample size tables for clinical studies*. 3rd ed.: Wiley-Blackwell.
318. Lewis, J.A. (1999). Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in Medicine*. **18**(15): p. 1903-1942.
319. Rochon, J. (1991). Sample size calculations for two-group repeated-measures experiments. *Biometrics*. p. 1383-1398.
320. Zhang, S. and C. Ahn. (2010). Effects of correlation and missing data on sample size estimation in longitudinal clinical trials. *Pharmaceutical Statistics*. **9**(1): p. 2-9.
321. Dawson, J.D. (1998). Sample size calculations based on slopes and other summary statistics. *Biometrics*. **54**(1): p. 323-330.
322. Galbraith, S. and I.C. Marschner. (2002). Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials*. **23**(3): p. 257-273.
323. Jung, S.H. and C. Ahn. (2003). Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine*. **22**(8): p. 1305-1315.
324. Yi, Q. and T. Panzarella. (2002). Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. *Controlled Clinical Trials*. **23**(5): p. 481-496.
325. Liang, K.-Y. and S.L. Zeger. (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B*. p. 134-148.
326. Fitzmaurice, G.M., N.M. Laird, and J.H. Ware. (2012). *Applied longitudinal analysis*. Vol. 998. John Wiley & Sons.
327. Fitzmaurice, G. and G. Molenberghs. (2009). Advances in longitudinal data analysis: an historical perspective. *Longitudinal Data Analysis*. p. 3-30.
328. Hedeker, D. and R.D. Gibbons. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
329. Hu, F.B., et al. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*. **147**(7): p. 694-703.
330. Nelder, J.A. and R.J. Baker. (2006). *Generalized linear models*, in *Encyclopedia of Statistical Sciences*. Wiley Online Library.
331. Li, F., et al. (1998). Analysis of longitudinal data of repeated observations using generalized estimating equations methodology. *Measurement in Physical Education and Exercise Science*. **2**(2): p. 93-113.
332. Winer, B.J. (1971). *Statistical principles in experimental design*. 2nd ed. McGraw-Hill Series in Psychology. New York: McGraw-Hill.
333. Bock, R.D. (1985). *Multivariate statistical methods in behavioral research*. USA: Scientific Software, Inc.
334. Molenberghs, G. and G. Verbeke. (2005). *Models for discrete longitudinal data*. Springer Series in Statistics. New York: Springer.

335. Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in statistics-Simulation and computation*. **22**(4): p. 1079-1106.
336. Littel, R., et al. (1996). *SAS system for mixed models*. 2nd ed. Cary, NC: SAS Institute.
337. Guerin, L. and W.W. Stroup. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data, in *Annual Conference on Applied Statistics in Agriculture.*
338. Diggle, P.J. (1988). An approach to the analysis of repeated measurements. *Biometrics*. **44**(4): p. 959-971.
339. Gibbons, R.D., D. Hedeker, and S. Dutoit. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*. **6**: p. 79-107.
340. Gibbons, R.D., et al. (1988). Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin*. **24**(3): p. 438-443.
341. De Leeuw, J. and I. Kreft. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*. **11**(1): p. 57-85.
342. Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*. **74**(4): p. 817-827.
343. Liu, C., et al. (2007). RANDOM and REPEATED statements: how to use them to model the covariance structure in Proc Mixed, in *SAS Conference Proceedings: Midwest SAS User Group.*
344. Johnson, M. (2002). Individual growth analysis using PROC MIXED. *SAS User Group International*.
345. Hamer, R. and P. Simpson. (2000). Mixed-up mixed models: things that look like they should work but don't, and things that look like they shouldn't work but do, in *Proceedings of the Twenty-Fifth Annual SAS® Users Group International Conference.*
346. Lord, F.M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*. **68**(5): p. 304-305.
347. Liu, G.F., et al. (2009). Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine*. **28**(20): p. 2509-2530.
348. Senn, S.J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*. **8**(4): p. 467-475.
349. Van Breukelen, G.J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*. **59**(9): p. 920-925.
350. Van Breukelen, G.J. (2013). ANCOVA Versus CHANGE From baseline in nonrandomized studies: the difference. *Multivariate Behavioral Research*. **48**(6): p. 895-922.
351. Lesaffre, E. and S. Senn. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine*. **22**(23): p. 3583-3596.
352. Raab, G.M., S. Day, and J. Sales. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*. **21**(4): p. 330-342.
353. Oakes, J.M. and H.A. Feldman. (2001). Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. *Evaluation Review*. **25**(1): p. 3-28.
354. Tu, Y.K., et al. (2005). Statistical power for analyses of changes in randomized controlled trials. *Journal of Dental Research*. **84**(3): p. 283-287.
355. Van Breukelen, G.J. (2006). ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies [corrected]. *Journal of Clinical Epidemiology*. **59**(9): p. 920-925.
356. Wright, D.B. (2006). Comparing groups in a before–after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology*. **76**(3): p. 663-675.
357. Vickers, A.J. (2004). Statistical reanalysis of four recent randomized trials of acupuncture for pain using analysis of covariance. *The Clinical Journal of Pain*. **20**(5): p. 319-323.
358. Cribbie, R.A. and J. Jamieson. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online*. **9**(1): p. 37-55.
359. Cox, D.R. and P. Mccullagh. (1982). A biometrics invited paper with discussion. some aspects of analysis of covariance. *Biometrics*. p. 541-561.
360. Lu, K. (2010). On efficiency of constrained longitudinal data analysis versus longitudinal analysis of covariance. *Biometrics*. **66**(3): p. 891-896.

361. Coffman, C.J., D. Edelman, and R.F. Woolson. (2016). To condition or not condition? Analysing 'change'in longitudinal randomised controlled trials. *BMJ Open*. **6**(12): p. e013096.
362. Dimitrov, D.M. and P.D. Rumrill, Jr. (2003). Pretest-posttest designs and measurement of change. *Work*. **20**(2): p. 159-165.
363. Zhang, S., et al. (2014). Empirical comparison of four baseline covariate adjustment methods in analysis of continuous outcomes in randomized controlled trials. *Clinical Epidemiology*. **6**: p. 227-235.
364. Assmann, S.F., et al. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. **355**(9209): p. 1064-1069.
365. Sargent, J., K. Coupland, and R. Wilson. (2005). SAS Institute Inc. *SAS Online Documentation*. **9**(3): p. 2002-2005.
366. Campbell, D.T. and D.A. Kenny. (1999). *A primer on regression artifacts*. New York: The Guildford Press.
367. Wainer, H. and L.M. Brown. (2006). 28 three statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data. *Handbook of Statistics*. **26**: p. 893-918.
368. Lydersen, S. (2015). Statistical review: frequently given comments. *Annals of the Rheumatic Diseases*. **74**(2): p. 323-325.
369. Glymour, M.M., et al. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemiology*. **162**(3): p. 267-278.
370. Winkens, B., et al. (2007). Randomized clinical trials with a pre- and a post-treatment measurement: repeated measures versus ANCOVA models. *Contemporary Clinical Trials*. **28**(6): p. 713-719.
371. Tu, Y.K., V. Baelum, and M.S. Gilthorpe. (2008). A structural equation modelling approach to the analysis of change. *European Journal of Oral Sciences*. **116**(4): p. 291-296.
372. (2001). *ICH Harmonised Tripartite Guideline: Choice of control group and related issues in clinical trials E10*. European Medicines Agency.
373. Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *International Journal of Psychophysiology*. **52**(3): p. 277-283.
374. Harrell, F.E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. 2nd ed. Springer Series in Statistics. Switzerland: Springer International Publishing.
375. Littell, R.C., et al. (2006). *SAS for mixed models*. 2nd ed. Cary, NC: SAS institute.
376. Littell, R., P. Henry, and C. Ammerman. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*. **76**(4): p. 1216-1231.
377. Littell, R.C., J. Pendergast, and R. Natarajan. (2000). Tutorial in biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*. **19**(1793): p. 1819.
378. Fleminger, S. (1994). Statistical notes. Regression towards the mean. *The BMJ*. **309**(6953): p. 539.
379. Bland, J.M. and D.G. Altman. (1994). Regression towards the mean. *The BMJ*. **308**(6942): p. 1499.
380. Littell, R.C., W.W. Stroup, and R.J. Freund. (2002). *SAS for linear models*. 4th ed. Cary, NC: SAS institute.
381. Muller, K.E. and B.A. Fetterman. (2002). *Regression and ANOVA: an integrated approach using SAS software*. Cary, NC: SAS Institute.
382. Vickers, A.J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*. **3**(1): p. 1.
383. Tu, Y.-K. and M.S. Gilthorpe. (2011). *Statistical thinking in epidemiology*. Boca Raton, FL: CRC Press.
384. Oldham, P.D. (1962). A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases*. **15**(10): p. 969-977.
385. Senn, S. (1991). Baseline comparisons in randomized clinical trials. *Statistics in Medicine*. **10**(7): p. 1157-1159.
386. Hertz, D.L. and H.L. Mcleod. (2013). Use of pharmacogenetics for predicting cancer prognosis and treatment exposure, response and toxicity. *Journal of Human Genetics*. **58**(6): p. 346-352.

387. Gillis, N.K., J.N. Patel, and F. Innocenti. (2014). Clinical implementation of germ line cancer pharmacogenetic variants during the next-generation sequencing era. *Clinical Pharmacology and Therapeutics*. **95**(3): p. 269-280.
388. Lewis, C.M. (2002). Genetic association studies: design, analysis and interpretation. *Briefings in Bioinformatics*. **3**(2): p. 146-153.
389. Bush, W.S. and J.H. Moore. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*. **8**(12): p. e1002822.
390. Lettre, G., C. Lange, and J.N. Hirschhorn. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*. **31**(4): p. 358-362.
391. Med.D.R.A. (2010). *Introductory guide MedDRA version 17.1*. Chantilly, VA: MedDRA Maintenance and Support Services Organization.
392. Liu, X. (2012). *Survival analysis: models and applications*. Chichester: John Wiley & Sons.
393. Du Bois, D. and E. Du Bois. (1989). A formula to estimate the approximate surface area if height and weight be known. 1916. *Nutrition*. **5**(5): p. 303.
394. Mosteller, R.D. (1987). Simplified calculation of body-surface area. *New England Journal of Medicine*. **317**(17): p. 1098.
395. Field, K.M., et al. (2008). Chemotherapy dosing strategies in the obese, elderly, and thin patient: results of a nationwide survey. *Journal of Oncology Practice*. **4**(3): p. 108-113.
396. Coate, L., et al. (2010). Germline genetic variation, cancer outcome, and pharmacogenetics. *Journal of Clinical Oncology*. **28**(26): p. 4029-4037.
397. Schork, N.J. (2015). Personalized medicine: Time for one-person trials. *Nature*. **520**(7549): p. 609-611.
398. Lawrence Gould, A., et al. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*. **34**(14): p. 2181-2195.
399. Mattila, K., et al. (1988). Blood pressure and five year survival in the very old. *British Medical Journal (Clinical Research Ed.)*. **296**(6626): p. 887-889.
400. Boutouyrie, P., et al. (2002). Aortic stiffness is an independent predictor of primary coronary events in hypertensive patients a longitudinal study. *Hypertension*. **39**(1): p. 10-15.
401. Tsiatis, A., V. Degruttola, and M. Wulfsohn. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. **90**(429): p. 27-37.
402. Wikby, A., et al. (1998). Changes in CD8 and CD4 lymphocyte subsets, T cell proliferation responses and non-survival in the very old: the Swedish longitudinal OCTO-immune study. *Mechanisms of Ageing and Development*. **102**(2): p. 187-198.
403. Wulfsohn, M.S. and A.A. Tsiatis. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*. **53**(1): p. 330-339.
404. Kwok, O.-M., S.G. West, and S.B. Green. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*. **42**(3): p. 557-592.
405. Ferron, J., R. Dailey, and Q. Yi. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*. **37**(3): p. 379-403.
406. Murphy, D.L. and K.A. Pituch. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*. **77**(3): p. 255-284.
407. Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*. **64**(5): p. 402-406.
408. Little, R.J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. **83**(404): p. 1198-1202.
409. Enders, C.K. (2010). *Applied missing data analysis*. New Yrok: The Guilford Press.
410. Little, R.J. and D.B. Rubin. (2014). *Statistical analysis with missing data*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.
411. Brown, H. and R. Prescott. (2015). *Applied mixed models in medicine*. 3rd ed. Chichester: John Wiley & Sons.
412. Moser, E.B. (2004). Repeated measures modeling with PROC MIXED, in *Proceedings of the 29th SAS Users Group International Conference.*

Appendices

Appendix A. Association between baseline characteristics and safety outcomes

Table 12.2.1 Univariate analysis of potential prognostic factors for neutropenia

Characteristic	*Ovarian*				*Lung*			
	Affected	*Unaffected*	*OR (95% CI)*	*p-value*	*Affected*	*Unaffected*	*OR (95% CI)*	*p-value*
Age	.	.	1.012 (0.989-1.036)	0.3151	.	.	1.028 (1.005-1.052)	0.0161
Sex	162	61		.	223	120		.
Female	162	61		.	94	53		.
Male	.	.		.	129	67	0.921 (0.589-1.441)	0.7192
ECOG Performance Status	162	61		.	223	120		.
0-1	110	48		.	200	99		.
2-3	30	8	0.611 (0.261-1.431)	0.2567	18	14	1.571 (0.751-3.289)	0.2306
Missing	22	5		.	5	7		.
Renal Function	162	61		.	223	120		.
0-1	139	48		.	191	96		.
2-3	17	11	1.874 (0.820-4.283)	0.1362	26	22	1.684 (0.907-3.125)	0.0988
Missing	6	2		.	6	2		.
Ethnicity	162	61		.	223	120		.
Caucasian	139	50		.	191	107		.
Asian	2	1	1.390 (0.123-15.665)	0.9499	2	4	3.570 (0.643-19.814)	0.1179
Black	12	6	1.390 (0.495-3.901)	0.9114	20	5	0.446 (0.163-1.223)	0.0721
Other	4	1	0.695 (0.076-6.368)	0.5204	6	2	0.595 (0.118-3.000)	0.3807
Mixed	1	1	2.778 (0.171-45.253)	0.5218	2	2	1.785 (0.248-12.854)	0.5728
Missing	4	2		.	2	.		.
Smoking	162	61		.	223	120		.
Never	45	19		.	25	16		.
Ex-smoker	11	3	0.646 (0.162-2.580)	0.6443	67	30	0.700 (0.327-1.498)	0.3822
Current	25	8	0.758 (0.290-1.979)	0.8351	120	68	0.885 (0.442-1.773)	0.8464
Unknown	81	31	0.906 (0.460-1.785)	0.6845	11	6	0.852 (0.263-2.763)	0.9997

Table 12.2.2 Univariate analysis of potential prognostic factors for gastrointestinal disorder

Characteristic	***Ovarian***				***Lung***			
	Affected	***Unaffected***	***OR (95% CI)***	***p-value***	***Affected***	***Unaffected***	***OR (95% CI)***	***p-value***
Age	.	.	1.020 (0.982-1.059)	0.3138	.	.	1.002 (0.972-1.033)	0.8931
Sex	203	20		.	298	45		.
Female	203	20		.	119	28		.
Male	.	.		.	179	17	0.404 (0.212-0.770)	0.0059
ECOG Performance Status	203	20		.	298	45		.
0-1	147	11		.	258	41		.
2-3	31	7	3.018 (1.084-8.401)	0.0344	28	4	0.899 (0.300-2.696)	0.8492
Missing	25	2		.	12	.		.
Renal Function	203	20		.	298	45		.
0-1	174	13		.	252	35		.
2-3	22	6	3.650 (1.259-10.580)	0.0171	39	9	1.662 (0.742-3.722)	0.2172
Missing	7	1		.	7	1		.
Ethnicity	203	20		.	298	45		.
Caucasian	172	17		.	258	40		.
Asian	2	1	5.059 (0.436-58.719)	0.9637	6	.	<0.001 (<0.001->999.999)	0.9764
Black	17	1	0.595 (0.075-4.752)	0.9877	21	4	1.229 (0.401-3.765)	0.9526
Other	4	1	2.529 (0.267-23.932)	0.9714	8	.	<0.001 (<0.001->999.999)	0.9732
Mixed	2	.	<0.001 (<0.001->999.999)	0.9762	3	1	2.150 (0.218-21.179)	0.9468
Missing	6	.		.	2	.		.
Smoking	203	20		.	298	45		.
Never	58	6		.	32	9		.
Ex-smoker	12	2	1.611 (0.289-8.968)	0.4273	86	11	0.455 (0.172-1.200)	0.2775
Current	31	2	0.624 (0.119-3.276)	0.4397	166	22	0.471 (0.199-1.117)	0.2564
Unknown	102	10	0.948 (0.328-2.741)	0.9127	14	3	0.762 (0.179-3.247)	0.7151

Appendix B. Manhattan plots

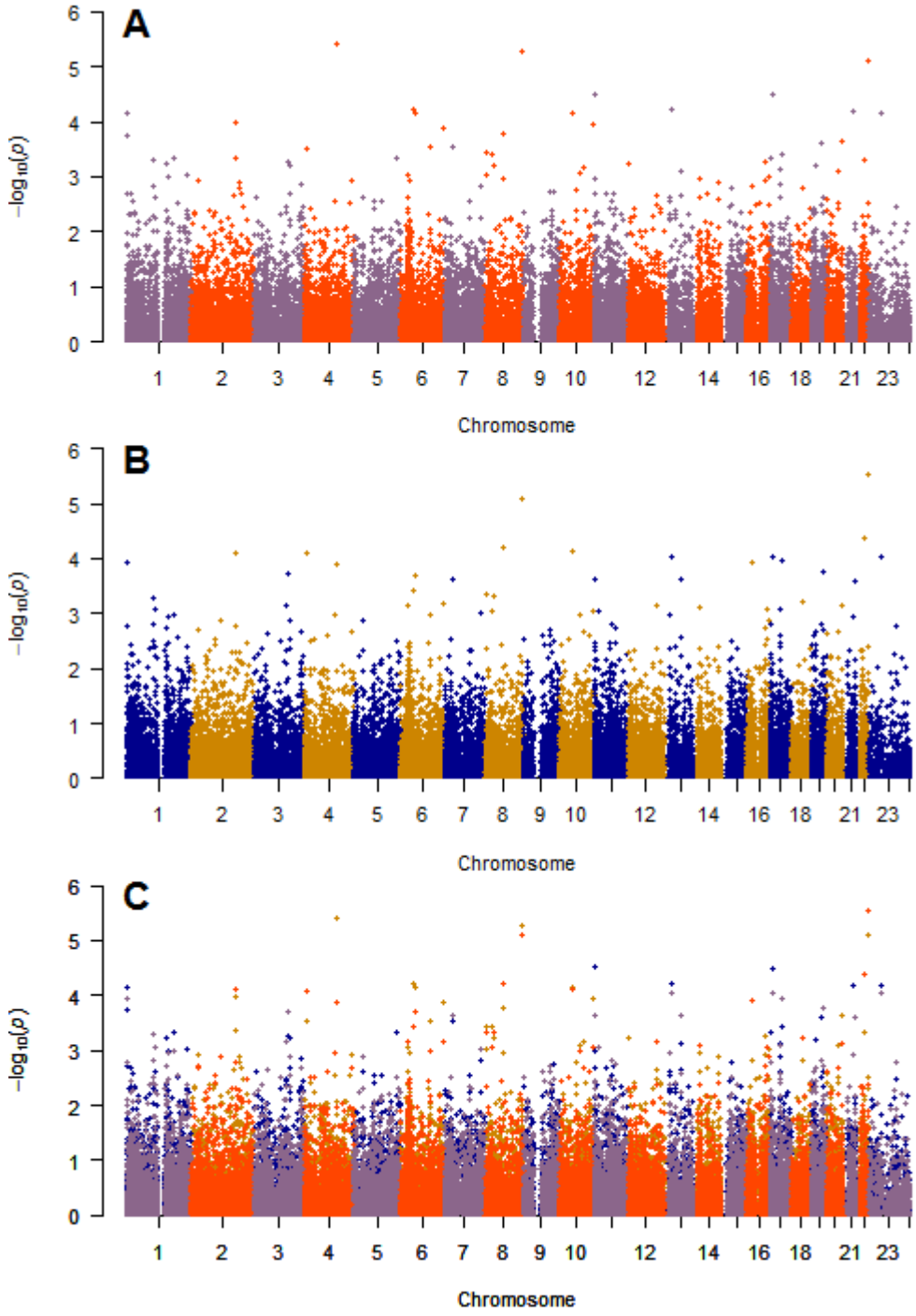


Figure 12.2.1 Manhattan plots of overall survival in the ovarian cohort.
The association of SNP genotype and overall survival was evaluated using a Cox regression models for 29328 SNPs across 223 patients with ovarian cancer treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

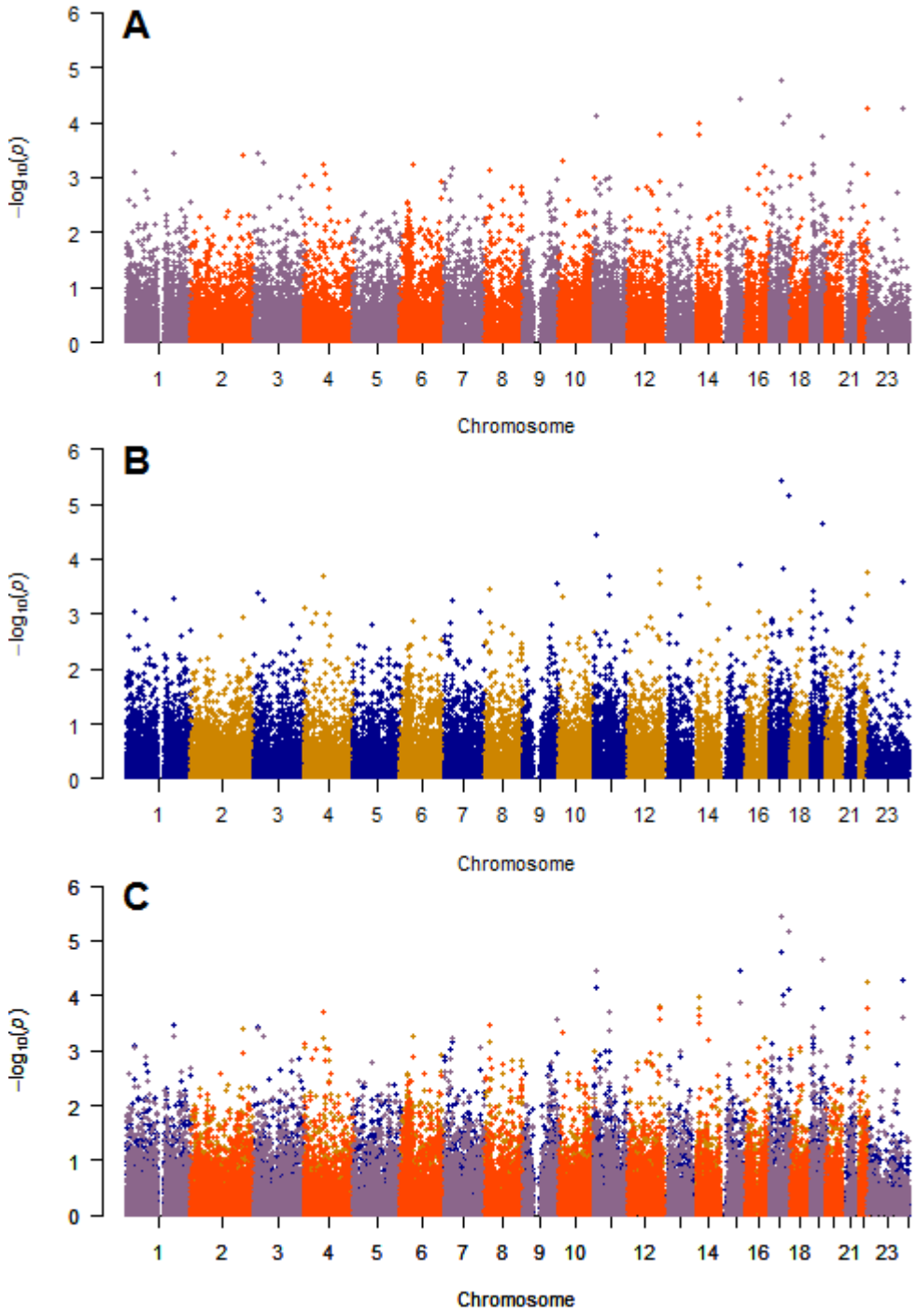


Figure 12.2.2 Manhattan plots of progression free survival in the ovarian cohort.
The association of SNP genotype and progression free survival was evaluated using a Cox regression models for 29328 SNPs across 223 patients with ovarian cancer uniformly treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

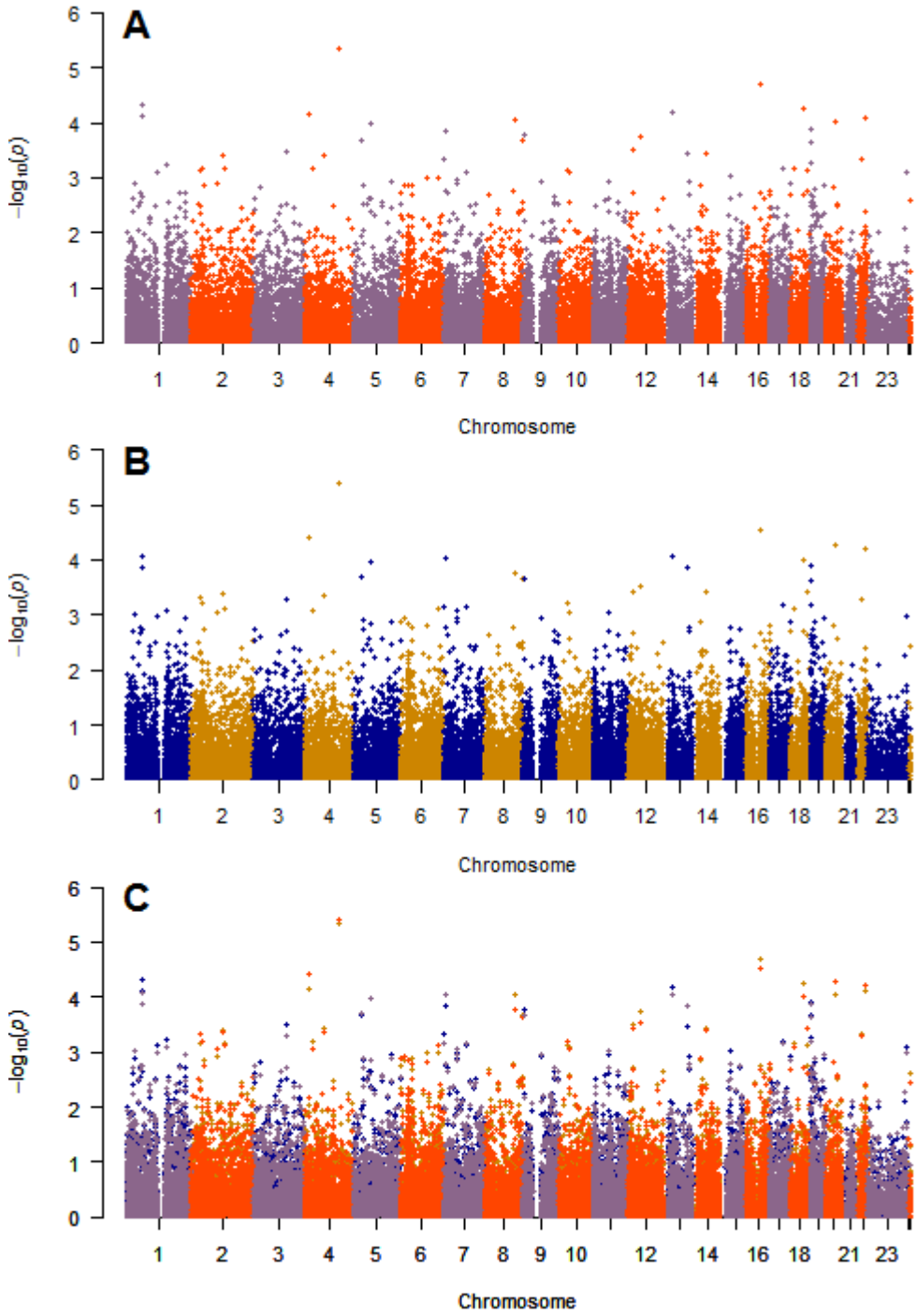


Figure 12.2.3 Manhattan plots of progression free survival in the lung cohort.
The association of SNP genotype and progression free survival was evaluated using a Cox regression models for 29328 SNPs across 343 patients with lung cancer treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

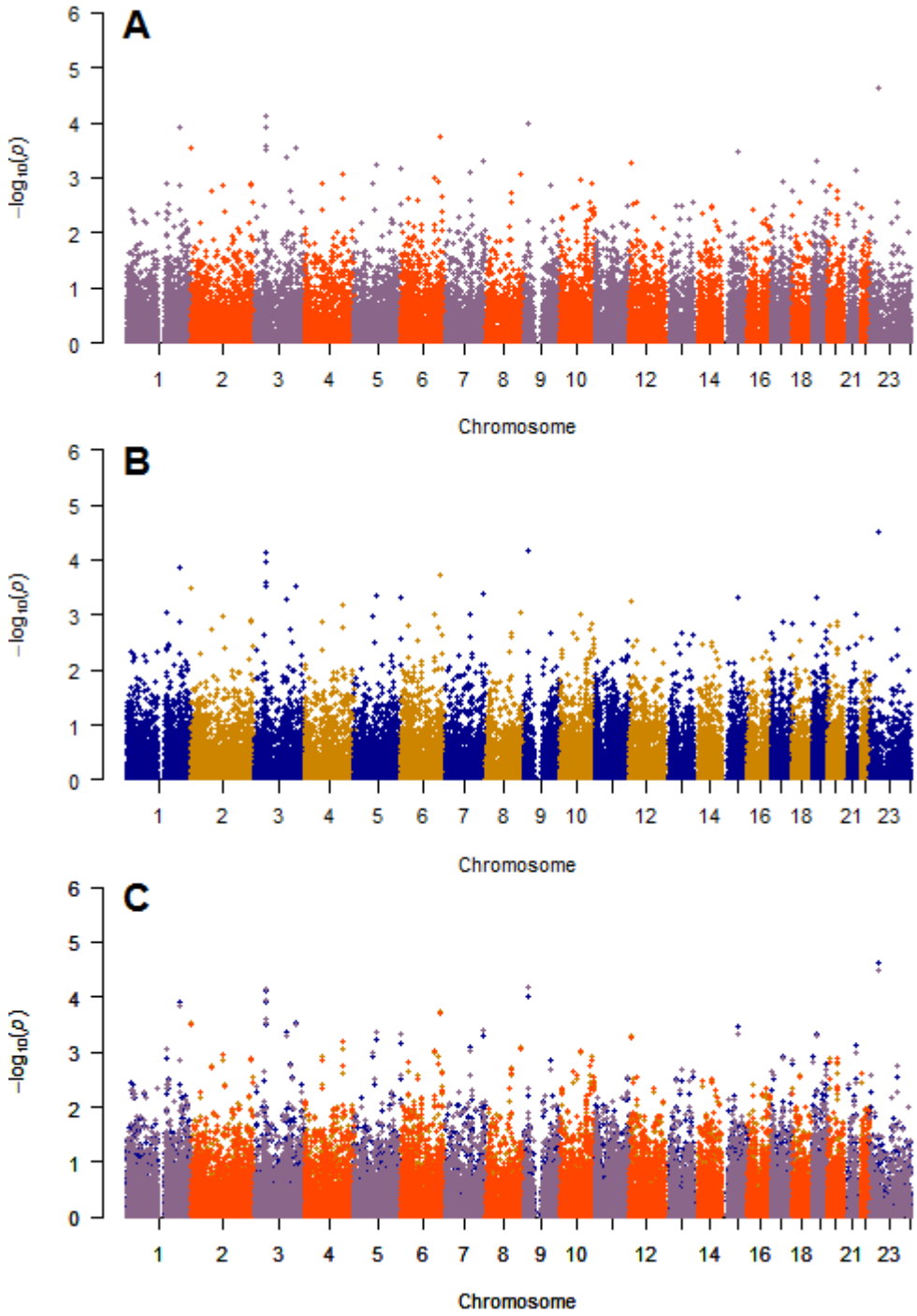


Figure 12.2.4 Manhattan plots of neutropenia in the ovarian cohort.
The association of SNP genotype and neutropenia was evaluated using a logistic regression models for 29328 SNPs across 223 patients with ovarian cancer treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

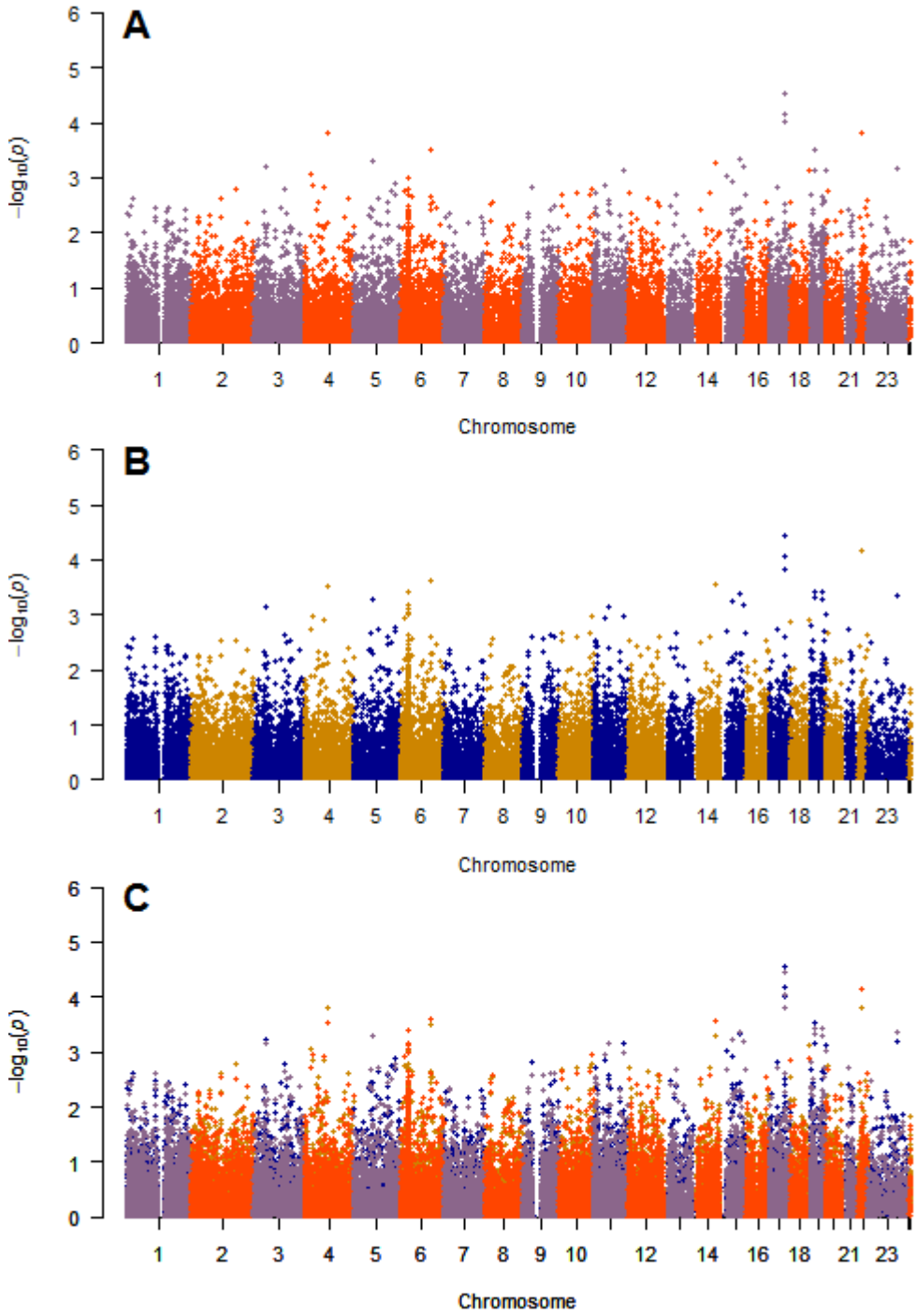


Figure 12.2.5 Manhattan plots of neutropenia in the lung cohort.
The association of SNP genotype and neutropenia was evaluated using a logistic regression models for 29328 SNPs across 343 patients with lung cancer treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

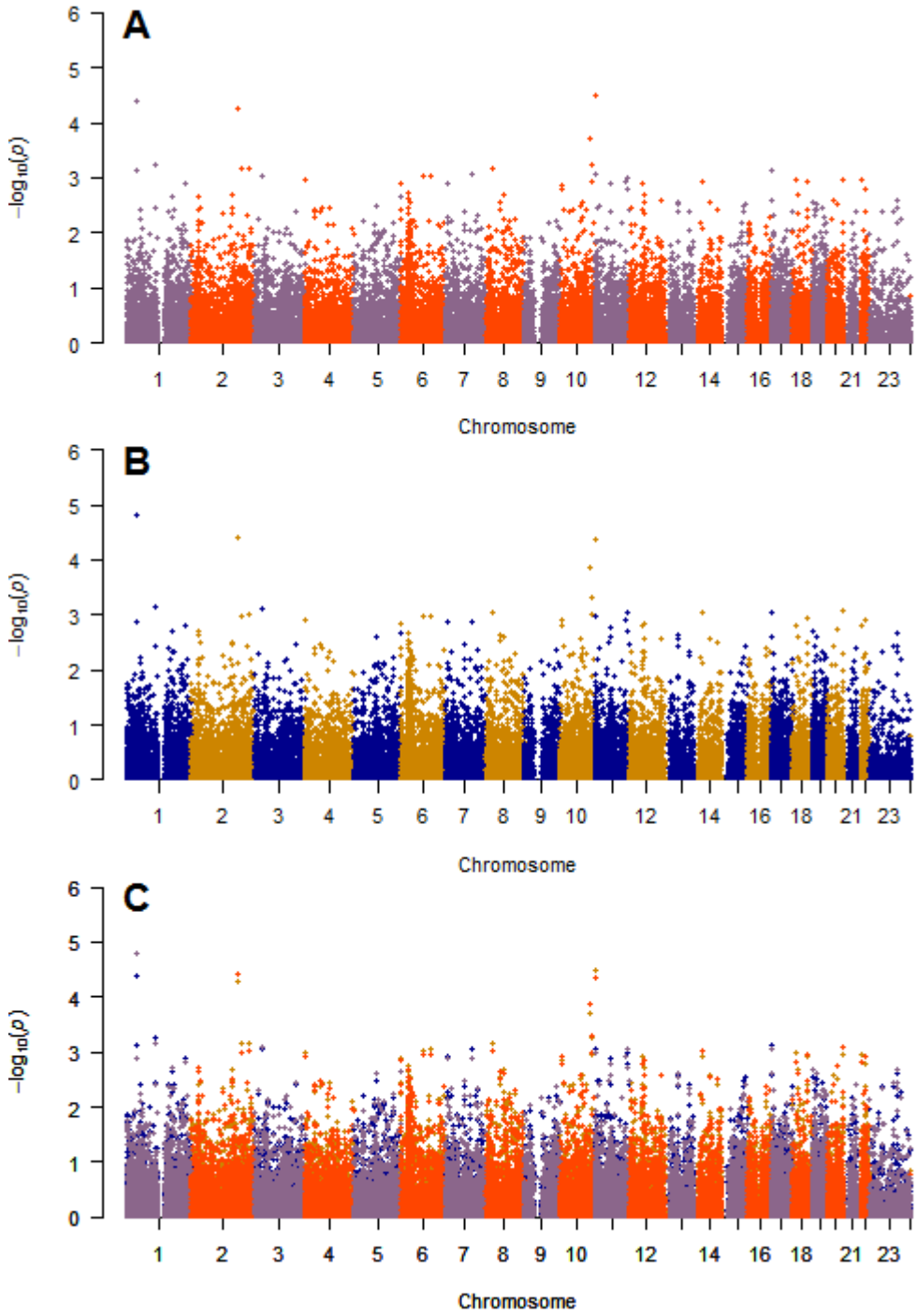


Figure 12.2.6 Manhattan plots of gastrointestinal disorder in the ovarian cohort.
The association of SNP genotype and gastrointestinal disorder was evaluated using a logistic regression models for 29328 SNPs across 223 patients with ovarian cancer treated with platinum therapy. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

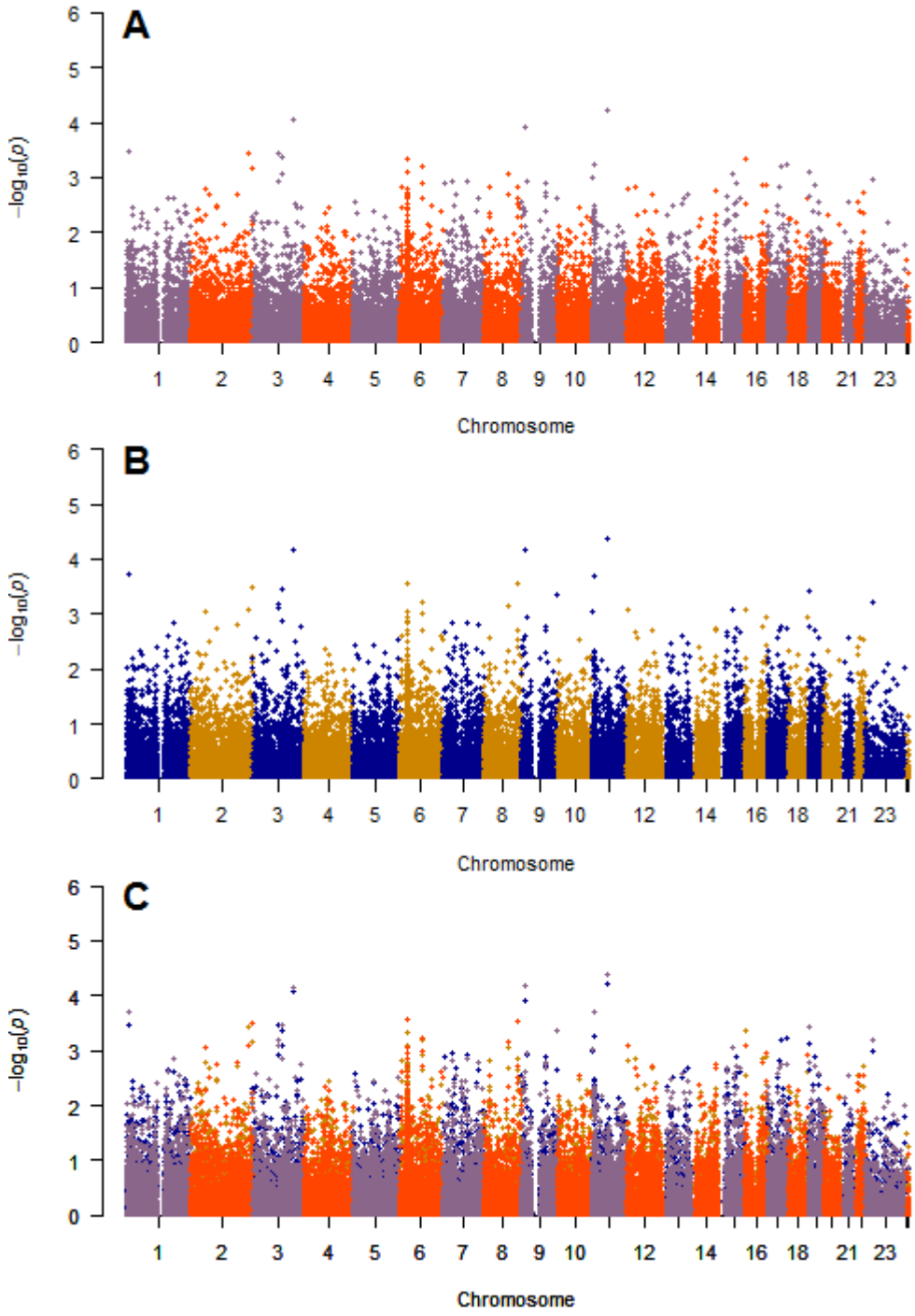


Figure 12.2.7 Manhattan plots of gastrointestinal disorder in the lung cohort.
The association of SNP genotype and gastrointestinal disorder was evaluated using a logistic regression models for 29328 SNPs across 343 patients with lung cancer treated with platinum-therapy containing regimens. Regression included *p*-values (−log10 *p*-values; y axis) are plotted against the respective chromosomal position of each SNP (x axis). (A) SNP Only results; (B) SNP Age Sex results; (C) Overlay plot of (A)-(B).

Appendix C. Quantile-Quantile (Q-Q) plots

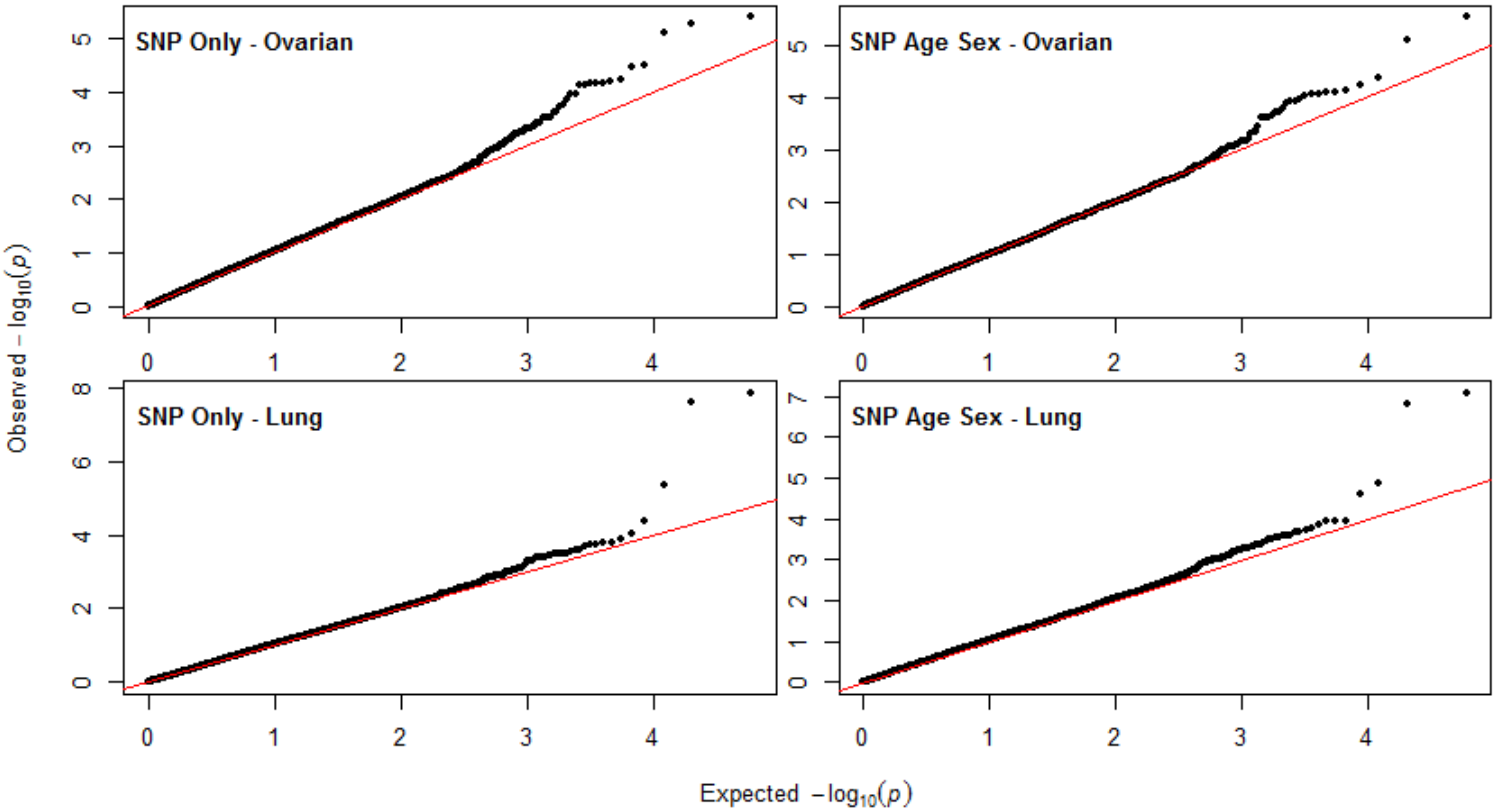


Figure 12.2.8 Q-Q-plots for overall survival results

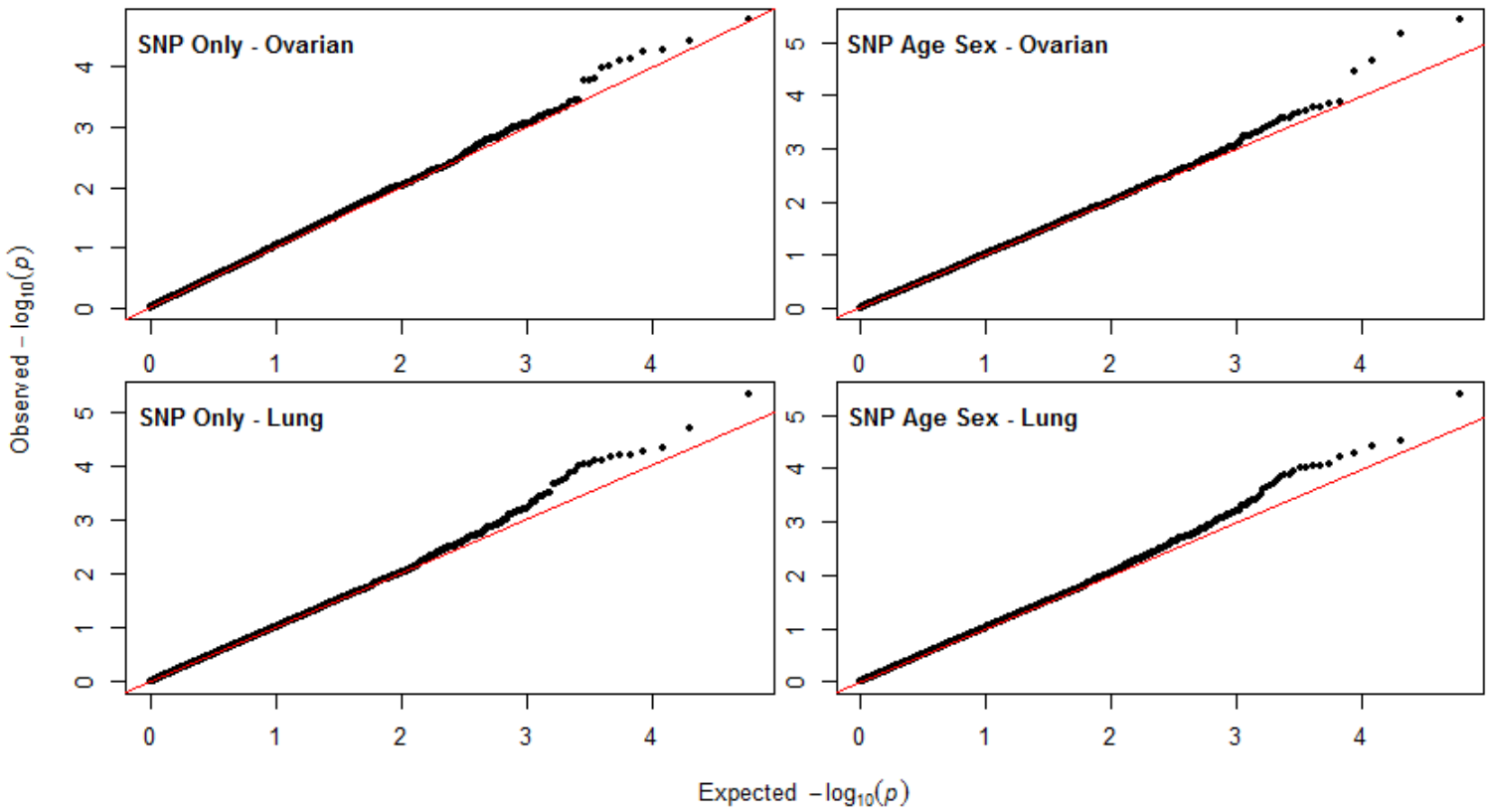


Figure 12.2.9 Q-Q-plots for progression free survival results

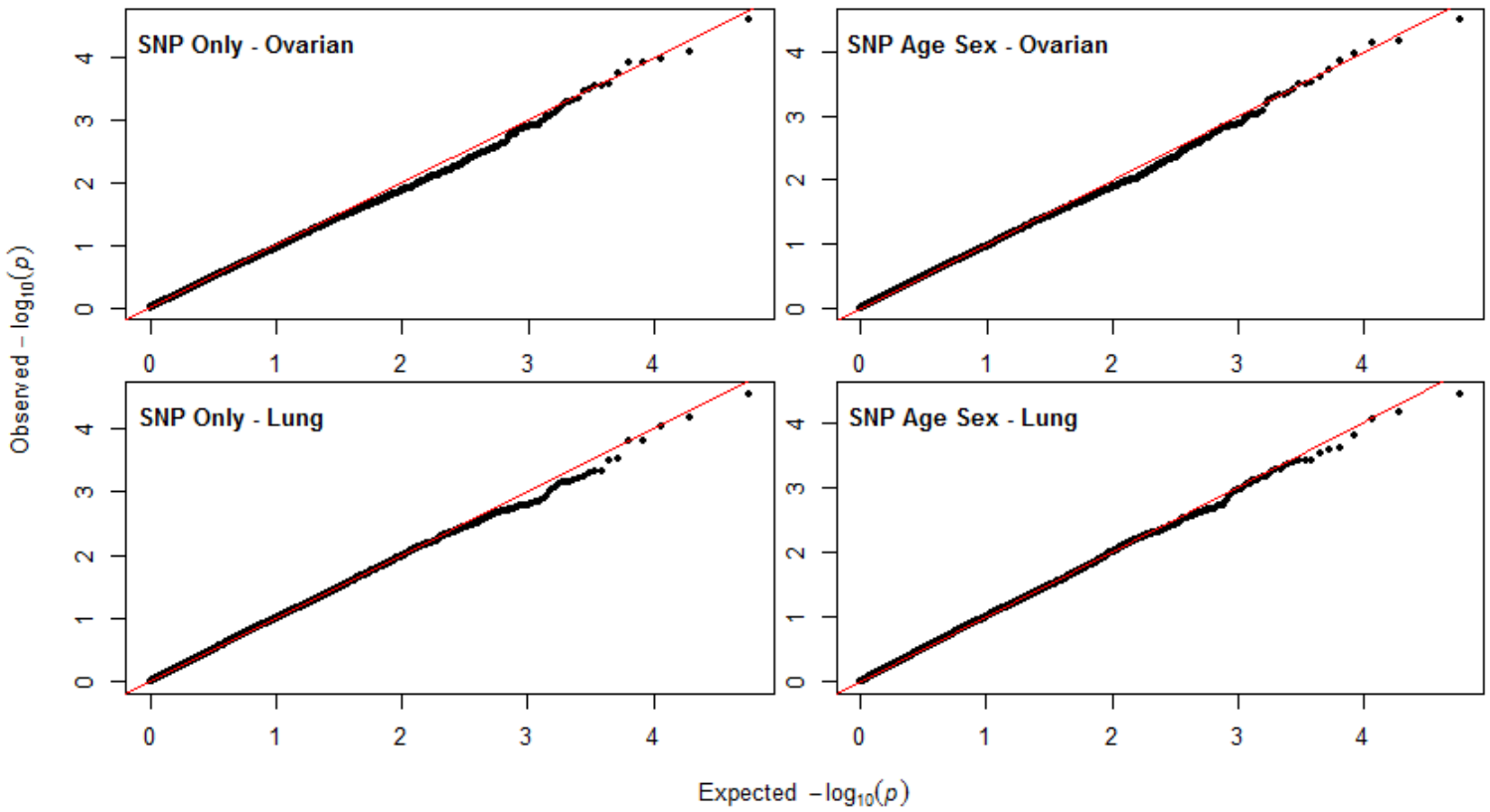


Figure 12.2.10 Q-Q-plots for neutropenia results

Appendix D. Power calculations using the sample size estimated by correlation structure misspecification

Table 12.2.3 Power under misspecification of the correlation structure.
Power for the Autoregressive(1) columns was calculated from the sample size required to provide 80% power with compound symmetry for the equivalent correlation strength and number of assessment time points. Power for the Compound Symmetry columns was calculated from the sample size required to provide 80% power with autoregressive(1) for the equivalent correlation strength and number of assessment time points.

Number of assessment time points	Δ	Autoregressive(1)				Compound Symmetry			
		Correlation Strength							
		0	0.25	0.5	0.75	0	0.25	0.5	0.75
2	0.1	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.25	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.5	0.80	0.81	0.81	0.81	0.80	0.81	0.81	0.81
3	0.1	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.25	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.5	0.80	0.81	0.81	0.81	0.80	0.81	0.81	0.81
4	0.1	0.80	0.76	0.76	0.76	0.80	0.85	0.85	0.85
	0.25	0.80	0.76	0.76	0.76	0.80	0.85	0.85	0.85
	0.5	0.80	0.76	0.77	0.78	0.80	0.85	0.86	0.87
5	0.1	0.80	0.71	0.71	0.71	0.80	0.89	0.89	0.89
	0.25	0.80	0.71	0.71	0.71	0.80	0.89	0.89	0.89
	0.5	0.81	0.71	0.72	0.72	0.81	0.90	0.89	0.90
6	0.1	0.80	0.66	0.66	0.66	0.80	0.93	0.93	0.93
	0.25	0.80	0.66	0.66	0.66	0.80	0.93	0.93	0.93
	0.5	0.80	0.66	0.67	0.69	0.80	0.93	0.93	0.93

Appendix E. Implementation of each baseline analysis method in SAS

P**roc mixed** is a SAS procedure based on mixed model methodology used to analyse longitudinal data [343, 412]. The syntax used to fit a linear model with fixed effects only in SAS is shown below, followed by a brief description of the primary statements:

```
6)  proc mixed data=mydata;
7)  class group subjid;
8)  model response = group time group*time / solution;
9)  repeated time / subject = subjid type=AR(1);
10) run;
```

The **proc mixed** statement calls the MIXED procedure. The **class** statement specifies which variables are classification (i.e. nominal) variables. The **model** statement is used to indicate the dependent (response) variable and to specify the fixed effects of the model. Our model includes fixed effects of *group*, *time*, and the interaction term *group*, *time.*

By specifying the **solution** option on the **model** statement we request *t*-tests and standard errors for each fixed effect (output into a table called “Solutionfor Fixed Effects”).

P**roc mixed** uses the `repeated` statement to model within subject variation. Here the repeated statement indicates that each subject (identified through their unique subjid has measurements repeated across the variable time. All subjects are assumed to have the same measurement covariance structure specified through the type=AR(1) code.

From a programming perspective, the difference between each of the three baseline analysis strategies is a function of 1) restricting the data to post baseline measurements only for the ANC and CSA analyses and 2) changing the model statement to match the scalar models.

For the RMA analysis, both the pre-post and multi-follow-up analyses use the same code:

```
11) proc mixed data=mydata;
12) class group subjid;
13) model response = group time group*time / solution;
14) repeated time / subject = subjid type=AR(1);
15) run;
```

Assuming that `mydata` is a dataset of pre-post design then the ANC analysis would adapt this code as follows:

```
16) proc mixed data=mydata(where=(time^=0);
17) class group subjid;
18) model response = group baseline/ solution;
19) repeated time / subject = subjid type=AR(1);
20) run;
```

The `(where=(time^=0)` restricts the input dataset to post baseline rows only and the model statement no longer includes time or its interaction with treatment. The variable baseline is a column containing the baseline response measurement for each patient.

If instead `mydata` was a multi-follow-up dataset then the ANC analysis code would take the following form:

```
21) proc mixed data=mydata(where=(time^=0);
22) class group subjid;
23) model response = group time group*time baseline/ solution;
24) repeated time / subject = subjid type=AR(1);
25) run;
```

For the CSA analysis, if `mydata` is a dataset of pre-post design then we would use the following code for analysis:

```
26) proc mixed data=mydata(where=(time^=0);
27) class group subjid;
28) model Delta_Resp = group / solution;
29) repeated time / subject = subjid type=AR(1);
30) run;
```

Where `Delta_Resp` is a change score calculated from `response - baseline`

And if `mydata` is a dataset of multi-follow-up design then the following code would be used for CSA analysis:

```
31) proc mixed data=mydata(where=(time^=0);
32) class group subjid;
33) model Delta_Resp = group time group*time / solution;
34) repeated time / subject = subjid type=AR(1);
35) run;
```

For each set of model parameters, data were simulated for 10,000 individuals using the simulation steps described above. Each data set was analysed using the SAS code, for each analysis

method (RMA, ANC, CSA) and model parameters were extracted and compared across the analysis methods.

Appendix F. Calculation of the average intra-patient correlation

Given the following example correlation matrix:

		Time point l			
		0	**1**	**2**	**3**
Time point k	**0**	**1**	**0.63**	**0.40**	**0.25**
	1	**0.63**	**1**	**0.63**	**0.40**
	2	**0.40**	**0.63**	**1**	**0.63**
	3	**0.25**	**0.40**	**0.63**	**1**

The average correlation between time points is (any row or column from the correlation matrix excluding k=l):

$$\frac{1}{m-1}\sum_{k=0}^{T}\rho_{0k} = \frac{0.63 + 0.40 + 0.25}{3} = 0.427$$

Appendix G. Pre-Post Covariance matrix

For a pre-post design, the covariance matrix will be of dimensions 2x2:

$$\mathrm{R}_i = \sigma^2 \begin{bmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho_{01} \\ \rho_{10} & 1 \end{bmatrix} \tag{12.2.1}$$

Where $\rho_{01} = \rho_{01}$, the correlation between pre and post measurements.

For a 2x2 matrix the inverse calculated by:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{12.2.2}$$

Assuming that $\sigma^2 = 1$ then the covariance matrix can be expressed as:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = 1 \begin{bmatrix} 1 & \rho_{01} \\ \rho_{10} & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{12.2.3}$$

And the inverse of the pre-post covariance matrix is therefore:

$$\begin{aligned} \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} &= \frac{1}{(1 \times 1) - (\rho \times \rho)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \\ &= \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{1 - \rho^2} & \frac{-\rho}{1 - \rho^2} \\ \frac{-\rho}{1 - \rho^2} & \frac{1}{1 - \rho^2} \end{bmatrix} = \begin{bmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{bmatrix} \end{aligned} \tag{12.2.4}$$

And we can see that $u_{11} = \frac{1}{1-\rho^2}$ which matches the inverse of the variance reduction associated with ANCOVA [381].

The sum of all the elements of this covariance matrix is:

$$\begin{aligned} w &= \frac{1}{1-\rho^2} + \frac{1}{1-\rho^2} + \frac{-\rho}{1-\rho^2} + \frac{-\rho}{1-\rho^2} \\ &= \frac{2-2\rho}{1-\rho^2} \\ &= \frac{2(1-\rho)}{(1+\rho)(1-\rho)} \\ &= \frac{2}{(1+\rho)} \end{aligned} \tag{12.2.5}$$

The ratio of $\frac{u_{11}}{w}$ is therefore:

$$\begin{aligned} \frac{\frac{1}{1-\rho^2}}{\frac{2}{(1+\rho)}} &= \frac{1}{(1-\rho)(1+\rho)} \times \frac{(1+\rho)}{2} \\ &= \frac{1}{2(1-\rho)} \end{aligned} \tag{12.2.6}$$

Which matches the inverse of the variance adjustment associated with the use of change scores [297].

Appendix H. Pre-post $X^T R^{-1} X$

Here we present the steps involved in the calculation of the $SE(\beta)$ for a pre-post study. For the ANC method the inversion of the $X^T R^{-1} X$ matrix requires that an assumption be made about the baseline measurement $E(Y_0)$ for both the control and active treatment arms. This is because the $(E(Y_0))^2$ will always be positive even when $E(Y_0) = 0$. In the calculations presented below we have assumed that $E(Y_0) = 0$ for both the control and active treatment groups and therefore $(E(Y_0)|control)^2 = 1$ and $(E(Y_0)|active)^2 = 1$

Variable	CSA	ANC	RMA
Design matrix	$\begin{bmatrix} 1 & R_1 \\ \vdots & \vdots \\ 1 & R_N \end{bmatrix}_{N\times 2}$	$\begin{bmatrix} 1 & R_1 & y_{10} \\ \vdots & \vdots & \vdots \\ 1 & R_N & y_{N0} \end{bmatrix}_{N\times 3}$	$\begin{bmatrix} 1 & R_1 & T_0 & R_1T_0 \\ 1 & R_1 & T_1 & R_1T_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & R_N & T_m & R_NT_m \end{bmatrix}_{(N\cdot m)\times 4}$
X^T	$\begin{bmatrix} 1 & \cdots & 1 \\ R_1 & \cdots & R_N \end{bmatrix}$	$\begin{bmatrix} 1 & \cdots & 1 \\ R_i & & R_N \\ y_{i0} & & y_{N0} \end{bmatrix}$	$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ R_i & R_i & \cdots & R_N \\ T_0 & T_1 & \cdots & T_m \\ R_iT_0 & R_iT_0 & \cdots & R_NT_m \end{bmatrix}_{4\times(N\cdot m)}$
R^{-1}	$\begin{bmatrix} \left(\frac{u_{11}}{w}\right)_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\frac{u_{11}}{w}\right)_N \end{bmatrix}$	$\begin{bmatrix} (u_{11})_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (u_{11})_N \end{bmatrix}$	$\begin{bmatrix} \begin{pmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{pmatrix}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \begin{pmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{pmatrix}_N \end{bmatrix}$
w	$\sum_{k,l=0}^{1} u_{kl}$		

Variable	CSA	ANC	RMA
$X^T R^{-1} X$	$\begin{bmatrix} \frac{Nu_{11}}{w} & \frac{Nu_{11}}{2w} \\ \frac{Nu_{11}}{2w} & \frac{Nu_{11}}{2w} \end{bmatrix}$	$\begin{bmatrix} Nu_{11} & \frac{Nu_{11}}{2} & 0 \\ \frac{Nu_{11}}{2} & \frac{Nu_{11}}{2} & 0 \\ 0 & 0 & Nu_{11} \end{bmatrix}$	$\begin{bmatrix} Nw & \frac{Nw}{2} & \frac{Nw}{2} & \frac{Nw}{4} \\ \frac{Nw}{2} & \frac{Nw}{2} & \frac{Nw}{4} & \frac{Nw}{4} \\ \frac{Nw}{2} & \frac{Nw}{4} & Nu_{11} & \frac{Nu_{11}}{2} \\ \frac{Nw}{4} & \frac{Nw}{4} & \frac{Nu_{11}}{2} & \frac{Nu_{11}}{2} \end{bmatrix}$
$(X^T R^{-1} X)^{-1}$	$\begin{bmatrix} \frac{2}{N\frac{u_{11}}{w}} & \frac{2}{N\frac{u_{11}}{w}} \\ \frac{2}{N\frac{u_{11}}{w}} & \frac{4}{N\frac{u_{11}}{w}} \end{bmatrix}$	$\begin{bmatrix} \frac{2}{Nu_{11}} & \frac{-2}{Nu_{11}} & 0 \\ \frac{-2}{Nu_{11}} & \frac{4}{Nu_{11}} & 0 \\ 0 & 0 & \frac{1}{Nu_{11}} \end{bmatrix}$	$\begin{bmatrix} \frac{-2u_{11}}{N({u_{01}}^2 - {u_{11}}^2)} & \frac{2u_{11}}{N({u_{01}}^2 - {u_{11}}^2)} & \frac{2}{N(u_{01} - u_{11})} & \frac{-2}{N(u_{01} - u_{11})} \\ \frac{2u_{11}}{N({u_{01}}^2 - {u_{11}}^2)} & \frac{-4u_{11}}{N({u_{01}}^2 - {u_{11}}^2)} & \frac{-2}{N(u_{01} - u_{11})} & \frac{4}{N(u_{01} - u_{11})} \\ \frac{2}{N(u_{01} - u_{11})} & \frac{-2}{N(u_{01} - u_{11})} & \frac{-4}{N(u_{01} - u_{11})} & \frac{4}{N(u_{01} - u_{11})} \\ \frac{-2}{N(u_{01} - u_{11})} & \frac{4}{N(u_{01} - u_{11})} & \frac{4}{N(u_{01} - u_{11})} & \frac{-8}{N(u_{01} - u_{11})} \end{bmatrix}$
$SE(\beta)$ for treatment effect	$= \sqrt{\frac{4}{N\frac{u_{11}}{w}}}$	$= \sqrt{\frac{4}{Nu_{11}}}$	$= \sqrt{\frac{-8}{N(u_{01} - u_{11})}} = \sqrt{\frac{8}{N(u_{11} - u_{01})}}$

Appendix G shows that:

$$\frac{u_{11}}{w} = \frac{1}{2(1-\rho)} \qquad u_{11} = \frac{1}{(1-\rho^2)} \qquad u_{01} = \frac{-\rho}{1-\rho^2}$$

We can therefore rewrite the $SE(\beta)$ for treatment effect for each analysis method as:

$SE(\beta)$ for treatment effect		
CSA	ANC	RMA
$= \sqrt{\frac{4}{N\frac{u_{11}}{w}}}$ $= \sqrt{\frac{4}{N\frac{1}{2(1-\rho)}}}$ $= \sqrt{\frac{8(1-\rho)}{N}}$	$= \sqrt{\frac{4}{Nu_{11}}}$ $= \sqrt{\frac{4}{N\frac{1}{(1-\rho^2)}}}$ $= \sqrt{\frac{4(1-\rho^2)}{N}}$	$= \sqrt{\frac{-8}{N(u_{01}-u_{11})}} = \sqrt{\frac{8}{N(u_{11}-u_{01})}}$ $= \sqrt{\frac{8}{N\left(\frac{1}{(1-\rho^2)} - \frac{-\rho}{(1-\rho^2)}\right)}}$ $= \sqrt{\frac{8}{N\frac{(1+\rho)}{(1-\rho)(1+\rho)}}} = \sqrt{\frac{8}{N\frac{1}{(1-\rho)}}}$ $= \sqrt{\frac{8(1-\rho)}{N}}$

Which confirms the equality of the $SE(\beta)$ for the RMA and CSA methods in a pre-post design.